The University Of Sheffield.

Genomic origins of novel metabolic pathways: the case of C_4 photosynthesis in grasses

by

Matheus Enrique Bianconi

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

August 2018

Contents

Acknowledgements

First, and foremost, I would like to thank my primary supervisor, Dr. Pascal-Antoine Christin, for his generous support throughout these years. His never-ending motivation and excitement about science are definitely inspiring. I am very grateful for his commitment and permanent willingness to help, which greatly facilitated the development of this work.

I would like to thank my co-supervisor, Professor Colin Osborne, for valuable advice since my first days in Sheffield. I am also very grateful to Dr. Guillaume Besnard, for his support and for generously sharing data and ideas, which greatly enriched this work.

I am also particularly grateful to Dr. Jill Olofsson and Dr. Luke Dunning, who kindly and constantly advised me on bioinformatic- and lab-related work, with great willingness, and also to Dr. Marjorie Lundgren, for kindly advising me on plant care and phenotyping so many times. I would like to thank also Dr. Jan Hackel, who along with Dr. Guillaume Besnard generously introduced me to the study system that resulted in my second chapter.

This work is the result of a joint effort of many people who collaborated in several ways. I would like to thank Emanuela Samaritani and Anne-Lise Liabot for kindly supporting me with lab work and plant care; Russell Hall, who taught me the basics of molecular biology and bioinformatics; Dr. Maria Vorontsova, who kindly provided plant material; Dr. Oriane Hidalgo and Dr. llia Leitch for helping with the genome sizing; Heather Walker and Gemma Newsome for support with mass spectrometry; Anthony Mapaura, for great support during field trip; the National Herbarium of Mozambique for support with collecting permits and field trip planning; Dr. Victor Soria-Carrasco and Dr. Helen Hipperson for bioinformatic support. A special thanks to Rachel Tucker, Gavin Horsburgh and Celine Pagnier, and all Molecular Ecology Lab people for continuous support with the lab work.

I am thankful to the University of Sheffield for hosting me, and to the staff of the Department of Animal and Plant Sciences, particularly to Angie Doncaster, who kindly

supported me with all student-related arrangements from the very beginning, and also to Allison Blake, John Beresford and Andy Krupa for committed support in various occasions. I also thank the AWEC and Annexe staff for continuous support with the plant growth facilities, and the staff of the High Performance Computing facilities.

I am also thankful to my viva examiners, Professor Jon Slate and Dr. Alex Papadopulos, for their insights and valuable suggestions on this study.

It was a great pleasure to have shared these years as a PhD student with my friends Jose Moreno-Villena, Danny Wood, Teera Watcharamongkol and Lamiaa Munshi, who are lovely people and have been very supportive. I also thank the members of Christin lab Nick Moody, Chatchawal Phansopa, Kui Ting Lin and Sam Hibdige, and all my officemates and APS cohorts for their support and pleasant company. I am also grateful to all members of Osborne, Nosil and Nadeau labs, who provided me with many insights on ecology, physiology and genomics in our weekly lab meetings.

I would like to thank the friends I have made in Sheffield, particularly the 'little Brazil' in the Atlantic One, who made these years much enjoyable. A special thanks to my longtime friend Pedro Godoy, who is an incredibly generous and supportive friend.

Agradeço aos meus pais, Milton e Jacira, por terem me dado incentivo e todas as condições para que eu pudesse hoje estar aqui hoje, finalizando uma tese de doutorado. Mesmo longe, eles certamente tiveram um papel enorme nisso tudo. Agradeço também ao meu irmão Thiago e sua família, por todo o apoio e carinho ao longo desses anos fora do país.

I am immensely grateful to my wife, Lígia Bertolino, who stood by me every moment during these years. Producing a dissertation has been a very demanding effort, and required constant encouragement, and lots of patience and care. She unconditionally offered me all of that, so I am sincerely thankful to her.

Finally, I would like to thank the Brazilian Ministry of Science and Technology through the National Research Council (CNPq), that ultimately made all of this possible by funding my PhD.

Summary

The existence of traits of impressive complexity has always puzzled evolutionary biologists. Traits such as camera eyes, bacteria flagella and plant carnivory result from intricate interactions between multiple structural and metabolic components. Understanding how each of these components originated and evolved to lead to the emergence of such a network of interactions and the associated function is therefore a major scientific challenge. Darwin and Wallace provided probably the major contribution to the problem; natural selection operating over successive generations via slight modifications can produce complexity. Nonetheless, there is still a large gap between the macroevolutionary patterns that are observed and the genetic changes underlying them. Here I address the problem of complex trait origins using the C_4 photosynthetic metabolism as a study system. My comparative analyses of whole genome sequencing data of selected grass lineages showed that (1) enzymes of the C_4 cycle can evolve via a burst of amino acid substitutions concentrated in a relatively short period of time, followed by continued adaptive evolution and anatomical specializations, showing that a single C_4 origin can give rise to a variety of C_4 phenotypes; (2) gene duplication via dosage effects can be a mechanism to suddenly increase the expression levels of genes involved in the C_4 cycle; (3) adaptive mutations in components of the C_4 trait can evolve in isolation in distinct genetic pools, and later be combined in admixture events; and (4) in some cases such adaptive mutations might be swept across populations by means other than recursive recombination. Overall, the findings presented in this dissertation suggest that (i) the components required for a rudimentary C_4 cycle might be acquired in a relatively short period of time via large effect mutations on key genes, and (ii) genetic exchanges between divergent lineages can facilitate the assembly and optimization of a C_4 metabolism. In addition, the methods developed here to analyse the genomic origins of C_4 photosynthesis using low-coverage sequence data can be applied to other groups and other traits, potentially contributing to the advent of large-scale comparative genomic analyses to understand the evolutionary origins of complex adaptive traits.

Chapter 1.
General Introduction

1. General Introduction

1.1. Complex traits

The origins of complex biological features have long been a conundrum for mankind. During most of the history of human societies, this topic belonged to natural theology and philosophy (Mayr 1982). Darwin and Wallace definitively claimed this problem to the realm of science, offering the concept of natural selection to explain the origin of complex traits, such as the eye, via "numerous, successive, slight modifications" of initial, rudimentary structures or functions (Darwin 1859). During the last century, key developments from evolutionary genetics have provided the theoretical ground to explain how complexity can arise in biological systems from natural selection operating successively over existing variation. However, our understanding of the stepwise evolutionary trajectories underlying novel complex adaptations in natural populations is still incipient.

The evolution of novel traits generally occurs via the co-option of preexisting structures and/or genes (Panganiban 1997; True and Carroll 2002; McLennan 2008; Shubin et al. 2009; Martinson et al. 2017). Examples include crystallins, which are multifunctional proteins that were co-opted as lens proteins in the eye (Wistow 1993), and insect wings, which might have evolved from the gills of ancestral aquatic arthropods (Damen et al. 2002). This recycling process implies that novel structures or genetic pathways are not expected to originate *de novo*, so that transitions between character states require fewer changes than suggested by the complexity of the traits. The co-option of structure or genes therefore can facilitate the accessibility to a novel trait by decreasing the evolutionary gap to an evolutionary innovation.

One major question concerning the origin of novel adaptive traits is the time scale of evolutionary change. Darwin supposed that evolutionary changes take place slowly, over long periods and that the changes are imperceptible when analysed under short time intervals (Darwin 1859). A classical alternative to Darwin's "phyletic gradualism" emerged from palaeontological studies showing that some lineages have little morphological variation during long periods in the geological record, and that these periods are interleaved by short periods of massive changes (Eldredge and Gould 1972). Although the so-called "punctuated equilibrium" theory was highly criticized due to attempts of their authors to make it a universal principle (Gould and Eldrege 1977;

Eldredge et al. 1997), this idea received some empirical support (e.g. Elena et al. 1996), and, most importantly, raised important discussions about the timing of evolutionary change and the ecological and genetic factors influencing it (Milligan 1986; Møller and Pomiankowski 1993; Pennell et al. 2014).

Fossil series documenting evolutionary transitions provide strong evidence for the inference of the time scale and the trajectories of trait transformations (e.g. body size evolution in horses, MacFadden 1986; land-to-water transition in cetaceans, Thewissen et al. 2001). However, this approach is obviously limited to organisms that are prone to fossilization and is constrained by the abundance of the fossil record for the particular trait. When the fossil record is not dense enough, inferring the history of the trait of interest must rely exclusively on comparative analyses of extant species using a phylogenetic approach. Comprehensive description of the phenotype states of living species captured in a phylogenetic tree allows inferring ancestral states and potential evolutionary trajectories using parsimony or more sophisticated statistical methods (Cunningham et al. 1998). The method used to code the character states is however important, as treating complex traits as binary characters hides the fact that their multiple constituents can have slightly different evolutionary histories. Studying the trajectory of each of these components is an approach that has been providing numerous insights into the evolution of complex traits (Christin et al. 2010a; Plachetzki et al. 2010; Dunning et al. 2017).

Studies based on phenotype analysis bridge the gap between disparate states and provide hypotheses for the order of changes that led to the origin of a complex trait. However, the microevolutionary aspects underlying the transitions, including the genetic mechanisms and the population level dynamics of genetic change, are often less well understood. The current availability of unprecedented genetic information of non-model species opens new avenues for investigating the factors promoting the evolution of complex adaptations in some groups (Nadeau and Jiggins 2010; Stapley et al. 2010; Ekblom and Galindo 2011; Messer et al. 2016).

1.2. Genomic origins of complex traits

The evolutionary assembly of complex adaptations is ultimately made possible by mechanisms that produce genetic variability, since natural selection operates over existing variation within populations (Darwin 1859). Mutation is the fundamental

source of genetic variation. In the broad sense, mutations include any change in the DNA sequence, from point changes (e.g. nucleotide substitutions, insertions, deletions) to major genomic rearrangements (e.g. segmental duplications, chromosomal inversions and translocations). Some particular forms of genomic rearrangements, such as gene duplications, have recurrently facilitated the origin of evolutionary novelties (Ohno 1970; Taylor and Raes 2004).

New genes usually originate from preexisting ones via segmental duplications (Ohno 1970; True and Carroll 2002; Zhang 2003). After gene duplication, the daughter copies can undergo different fates (Lynch and Conery 2000; Taylor and Raes 2004; Conant and Wolfe 2008). Most duplicates accumulate silencing mutations under neutral processes or selection to reduce expression levels. This leads to frequent losses of duplicated genes, so that only a fraction of the duplicates are retained over time. On the other hand, the presence of duplicates can be beneficial from their birth. This is the case when increases in gene expression levels due to dosage effects are advantageous (e.g. Mouchès et al. 1986; Chen et al. 2008; Cook et al. 2012). Over longer evolutionary periods, duplicated genes that are retained due to its adaptive value or to neutral processes lead to functional redundancy. One of the two copies can then acquire a new function through the successive accumulation of beneficial amino acid substitutions (neofunctionalization), while the other copy retains the ancestral function. Alternatively, if the ancestral gene had multiple functions, the duplicated copies can accumulate substitutions independently and specialize into one of these functions (subfunctionalization). Such processes operate on large scales in cases of whole genome duplication, and can create a huge reservoir of genes which may be co-opted for novel functions during the evolutionary diversification of the group. Whole genome duplications might indeed have been key events to animal and plant evolutionary diversification (Burke et al. 1995; Hughes and Kaufman 2002; Adams and Wendel 2005; Otto 2007).

The fate of point mutations and major rearrangements is dictated by processes at the population level. The gene pool of a population results from the balance between selection and neutral processes, and includes variants that can potentially contribute to adaptive innovation. The accumulation of genetic variants depends on effective population sizes and population structure, in addition to the mutation rate (Hartl and Clark 2007). In particular, gene flow between divergent populations can increase genetic diversity and generate new genetic combinations that can facilitate the

acquisition of novel adaptations (Stern 2013). In the classical case, contact occurs between previously isolated populations of the same species (admixture), but genetic exchanges can also occur between different species (hybridization). During hybridization, gene variants that evolved separately can be combined, diversifying the substrate of natural selection. Lateral gene transfer (LGT), i.e. the transfer of genetic material across normal mating barriers (Keeling and Palmer 2008), potentially allows genetic exchanges among even more distantly related organisms, and may represent an additional source of innovation. This process has been a major driver of prokaryote evolution, so that a large fraction of the bacterial and archaeal genes were acquired via LGT (Ochman et al. 2000; Zhaxybayeva and Doolittle 2011; Nelson-Sathi et al. 2015). Interestingly, several recent studies have also reported adaptive LGT in eukaryotes (Hotopp et al. 2007; Keeling and Palmer 2008; Yoshida et al. 2010; Christin et al. 2012a; Li et al. 2014)

The identification of genomic rearrangements and their dynamics within and across species often relies on detailed genome-wide analyses. These in turn are facilitated by the existence of vast genetic resources, including complete genome models. Obtaining such resources has been greatly facilitated by the development of high-throughput methods, which have revolutionized the field of evolutionary genetics.

1.3. Genomics

One of the aims of evolutionary genetic studies is to link the proximal, mechanistic processes that determine the functions and structures of biological systems, to the ultimate, evolutionary drivers of such systems. Indeed some of the major contributions to our understanding of living systems came throughout the last century from the studies of inheritance and its molecular basis (Mayr 1982). In the early days of genetics, most of the insights into the basis of inheritance resulted from crossing experiments. The study of molecular polymorphisms using electrophoresis, which was later greatly benefited from the invention of the polymerase chain reaction (PCR), was consolidated as a major tool in genetics (Hartl and Clark 2007). DNA sequencing became possible only after the development of chain termination reaction (Sanger et al. 1977). The Sanger method became the prevalent sequencing technique for the subsequent 30 years; nonetheless, the high costs of reagents and instruments required for the technique prevented its widespread application in genetic studies.

The Sanger method provides DNA sequences up to ~ 1000 base pairs (bp) long, which is useful for studying single genes or short loci. Obtaining complete genome sequences with the method is therefore laborious, even though it was facilitated by later improvements on the method (e.g. shotgun sequencing; Anderson 1981). The first complete genome sequence of a free-living organism was 1.8 million bp (Mb) long and belonged to the bacterium *Haemophilus influenzae* (Fleischmann et al. 1995). Model organisms with even larger genomes were subsequently analyzed, with the first plant (*Arabidopsis thaliana*) being sequenced five years later (The Arabidopsis Genome Initiative 2000). The biggest achievement of the whole genome sequencing era using Sanger method was the human genome, the first vertebrate to be completely sequenced (Human Genome Sequencing Consortium 2001, 2004). A major initiative that started with the publication of the human genome then promoted the exponential development of faster and cheaper DNA sequencing technologies, which would culminate with next generation sequencing (NGS) technologies (Bentley et al. 2008; Ansorge 2009; Shendure and Ji 2008). NGS revolutionized molecular biology as it decreased ~ 10,000-fold the costs of whole genome sequencing in less than 10 years, making genomic studies of non-model species and large-scale population genomic analyses feasible even for small research groups (Stapley et al. 2010; van Dijk et al. 2014).

NGS has numerous applications, including whole-genome and transcriptome (RNA-seq) sequencing, as well as methylome analyses and restriction-site associated sequencing (RAD-seq). Each approach is better suited for the analysis of a given level of genome structure and function. Whole-genome sequencing of pooled samples is a particularly versatile approach to compare the genomes of multiple species or individuals (Buerkle and Gompert 2013; Schlötterer et al. 2014; Sims et al. 2014). Indeed, pooling samples before sequencing reduces the overall cost per sample, making such approaches affordable for a larger number of research groups. However, increasing the number of samples per sequencing batch comes at the expense of sequencing depth, leading to a low number of reads covering each site in the genome. This reduces the amount of genetic information per sample, potentially decreasing the accuracy of the genotyping effort. Nonetheless, such low-coverage approaches can be well suited for large-scale population genetic analyses where markers spread across the genome are needed for a large number of samples. Whole-genome sequencing at low coverage has also been particularly useful for phylogenetic studies, since a higher proportion of reads are obtained for organellar genomes, which are present in numerous copies within each

cell. These organellar genomes provide phylogenetically useful markers, and this approach has promoted the phylogenomics era, particularly in plants (Dodsworth 2015). However, restricting NGS to analyses of organellar genomes is somehow a waste of resources, since the nuclear information is also provided by the sequencing efforts. Therefore, instead of a few markers for a small number of species, NGS provide the opportunity to study high numbers of markers from the nuclear and organellar genomes for many species, so that many longstanding questions in evolutionary genetics can be tackled. In this sense, phylogenetic tools are especially well suited to infer evolutionary events from such vast genomic resources.

1.4. Phylogenetics and molecular evolution

A robust phylogenetic hypothesis is fundamental for addressing questions related to the evolution of any particular group, gene or trait. Establishing homology of characters, which denotes shared ancestry, is necessary before inferring phylogenetic trees. Indeed, the comparison of non-homologous characters can lead to spurious groupings and wrong phylogenetic hypotheses (Felsenstein 2003). Phylogenetic hypotheses based on morphological data, in particular, rely on the expertise of well-trained specialists to hypothesize that two characters in different specimens derive from the same character in their common ancestor. On the other hand, in phylogenetic trees based on molecular data (i.e. amino acid or nucleotide sequences), homology is inferred from the optimal alignment between sequences, which is calculated by sequence alignment algorithms. Possibly the main advantage of the molecular approach over morphology-based analyses is that homology inference becomes less subjective, therefore more reproducible. Analyses of gene and protein sequences remain however complicated by a number of processes that characterize molecular evolution. These include gene duplications, convergent evolution, incomplete lineage sorting and lateral gene transfer (LGT).

Gene duplication is a pervasive phenomenon, so that organisms generally carry multiple genes that derive from a single copy that was duplicated in their ancestor, and which have evolved along independent trajectories. These genes often retain some degree of similarity, so that only comprehensive comparative analyses can distinguish between genes that are related because they were duplicated in an ancestor lineage (i.e. paralogs), and genes that diverged from each other only via speciation events (i.e.

orthologs; Fig. 1.1). Identifying orthologs is therefore key for inferring phylogenetic relationships among species, as their history reflects the successive speciation events.

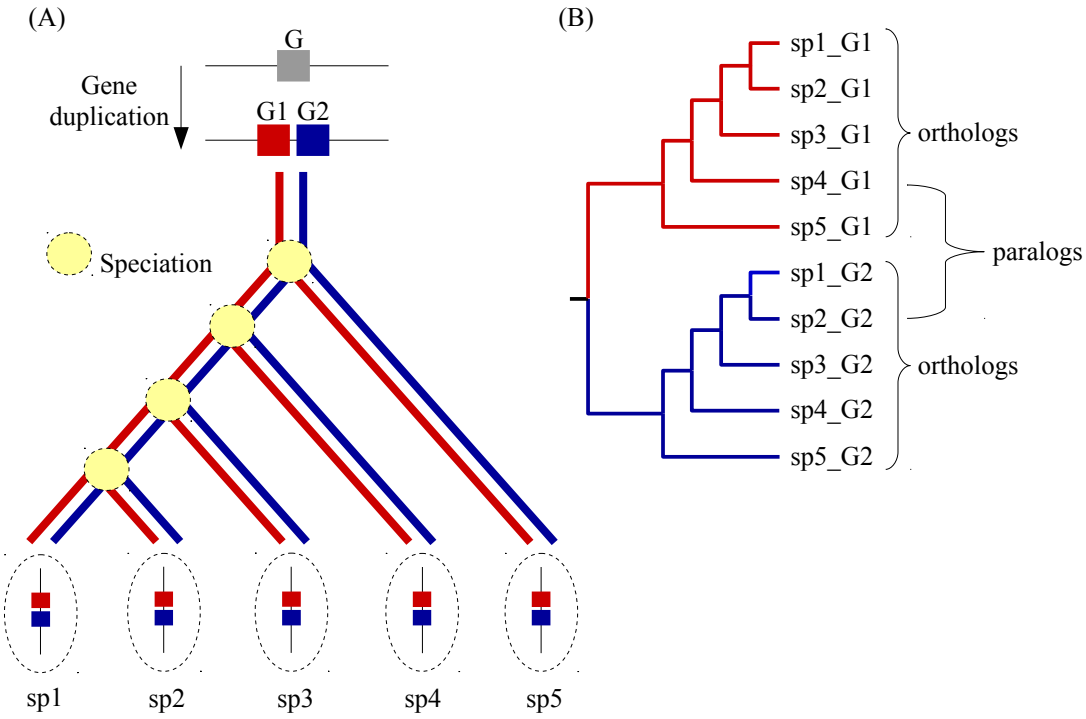


Fig. 1.1. Homology inference from gene trees. (A) Evolutionary trajectory of duplicated genes along four successive speciation events, and (B) the resulting gene tree, with orthology and paralogy relationships indicated.

Sequences of distantly related organisms that evolve under similar selective pressures can also confound phylogenetic inference. Indeed, sequences between distantly related organisms may share many sites that are not shared with closely related lineages that did not evolve under similar selective pressures. This phenomenon of molecular convergence has been reported in many groups and can be strong enough to lead to spurious groupings (Chen et al. 1997; Christin et al. 2007, 2010b; Fig. 1.2).

In cases of lineages that derive from a rapid event of speciation, different alleles of the same genes may not be completely sorted along the descendant lineages (Rosenberg 2003). As a consequence, some genes may be more similar between two divergent lineages than between individuals within the same lineage (Fig. 1.3). This issue, known as incomplete lineage sorting, is particularly relevant in within-species comparisons, or between groups that underwent successive speciation events separated by short periods of time.

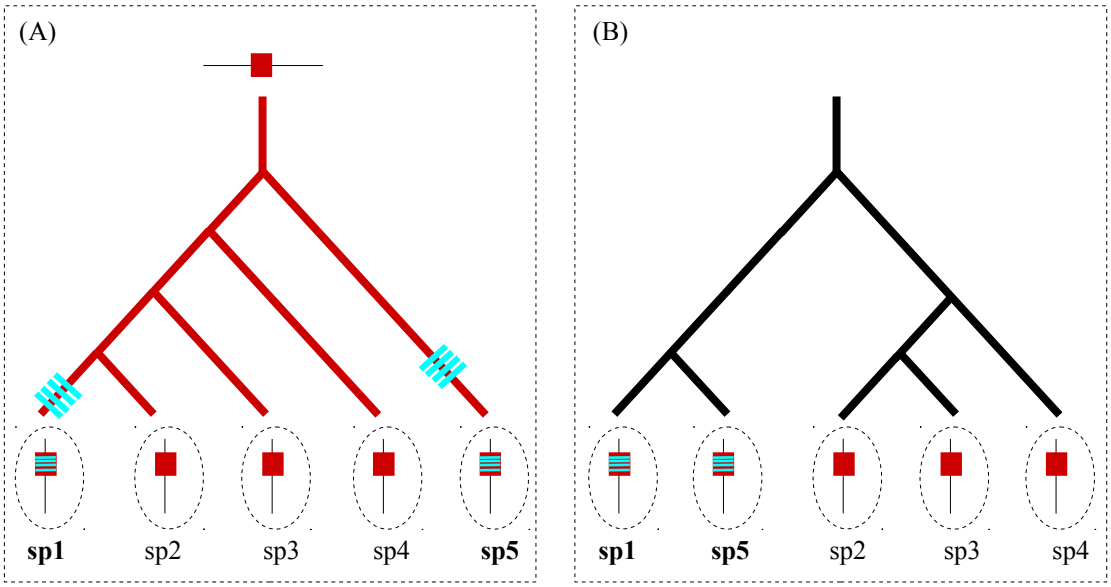


Fig. 1.2. Convergence at the molecular level. (A) Evolutionary trajectory of a gene that underwent identical substitutions in two independent lineages (sp1 and sp5), indicated as blue bars, and (B) an erroneous gene tree resulting from the grouping of distantly related genes with convergent substitutions.

The acquisition of genes from more distantly related lineages via LGT also causes mismatches between gene and species trees (Keeling and Palmer 2008). In this case, sequences acquired via LGT nest within the group of the donor lineage instead of within its correct group (Fig. 1.4). Such an issue is particularly relevant in prokaryotes where LGT is pervasive (Ochman et al. 2000; Keeling and Palmer 2008).

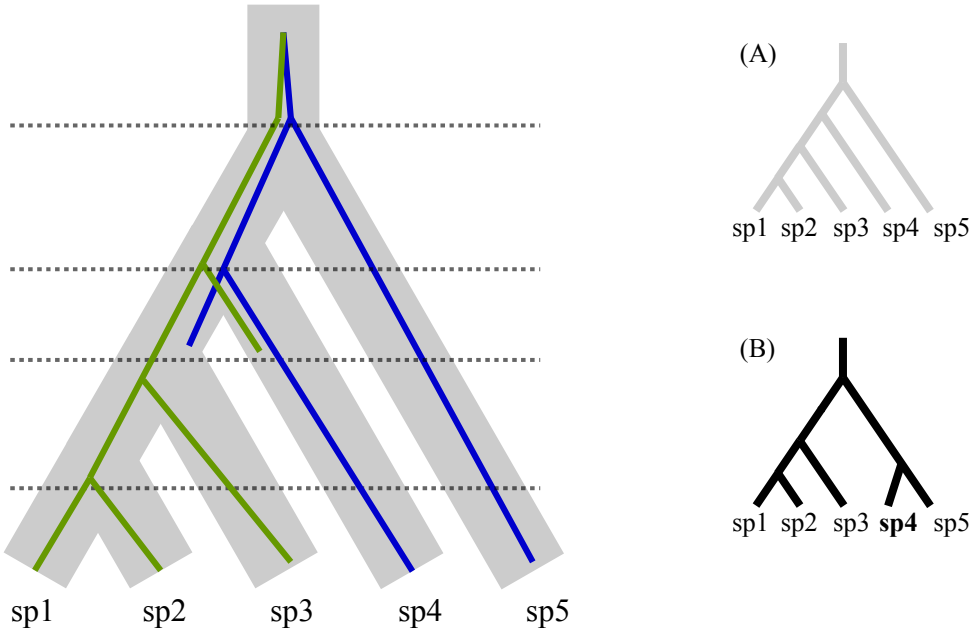


Fig. 1.3. A scenario of incomplete lineage sorting. Evolutionary trajectory of two genes (blue and green) along four successive speciation events (delimited by dashed lines). (A) Real relationship among species; (B) erroneous tree inferred from the blue and green genes.

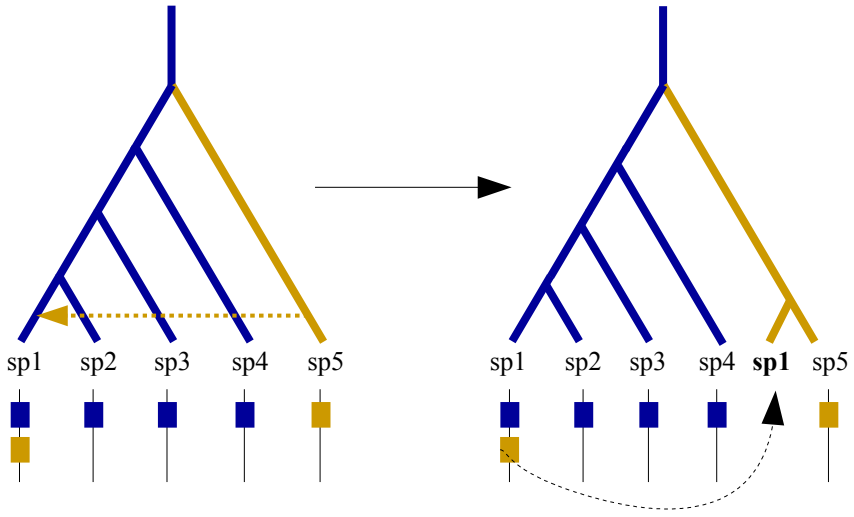


Fig. 1.4. Lateral gene transfer as evidenced by a gene tree. A gene that is transferred from 'sp5' to 'sp1' (left), and the expected gene tree (right).

While these phenomena can complicate phylogenetic inference, their detection provides important insights into the mechanisms of molecular evolution that characterize a particular gene, trait or group. For example, phylogenetic methods are major tools to detect the occurrence of LGTs (Keeling and Palmer 2008; Christin et al. 2012a; Mahelka et al. 2017) and convergent evolution at the molecular level (Chen et al. 1997; Christin et al. 2007; Rokas and Carroll 2008; Castoe et al. 2009). In addition, analyses of phylogenies for specific genes allows the inference of past selective pressures and reconstructing ancestral proteins, or even gene content in very distant ancestors.

Analysis of gene or protein sequences are also particularly important tools to estimate the timing of evolutionary events. Indeed, as substitutions accumulated over time, the number of differences between two sequences is proportional to their divergence time. Using genetic differences to estimate divergence times was first proposed by Zuckerkandl and Pauling (1965) and lead to the hypothesis of the molecular clock. Subsequent studies have shown that this assumption is not valid for most genes and lineages, since evolutionary rates vary as a function of intrinsic and extrinsic factors (Langley and Fitch 1974; Thorpe 1982; Felsenstein 2003; Kumar 2005). Models were therefore developed to account for this variation across lineages, and include the so-called relaxed clocks (e.g. Hasegawa and Kishino 1989; Thorne et al. 1998). Most of current methods of divergence times are performed on a Bayesian framework, with absolute dates obtained from fossils being used as priors to calibrate relaxed molecular clocks. Such approaches provide not only the divergence times

between lineages, but also a wide range of parameters of evolutionary significance, such as the rate of evolution and parameters related to speciation rates and trait evolution (Drummond and Rambaut 2007).

Overall, phylogenetics provide the framework to conduct comparative analyses, which rank amongst the most powerful tools to obtain information from past evolutionary events and species comparisons. Phylogenetic methods are especially important to study the evolutionary origins of traits of high complexity, either to infer the relationships among taxa and therefore the order of changes along the phylogeny, or to examine the mechanisms of molecular evolution involved in the evolutionary assembly of the trait. Instances of phenotypes that evolved independently in different lineages are particularly interesting study systems. Indeed, the multiple origins of a trait represent replicates of the evolutionary process, which increases our power to differentiate coincidence from causation. One of the most striking example of a complex phenotype that repeatedly evolved in eukaryotes is the C_4 photosynthetic metabolism, with 62 independent origins in land plants (Sage et al. 2011).

1.5. C_4 photosynthesis as an example of a complex trait

Photosynthetic organisms assimilate atmospheric CO_2 via a set of reactions that convert light into chemically available energy as ATP and NADPH, which in turn are consumed to reduce CO_2 to sugars through a series of energy-dependent reactions known as Calvin-Benson cycle, or photosynthetic carbon reduction (PCR). Both light reactions and the PCR cycle first evolved ~ 2.4 billion years ago, and are common to all organisms performing oxygenic photosynthesis (Hohmann-Marriott and Blankenship 2011). The primary enzyme catalysing carbon fixation via PCR cycle is ribulose 1,5-bisphosphate carboxylase/oxygenase (Rubisco; Taiz and Zeiger 2010). Besides its capacity to fix CO_2 in the 3-carbon compound RuBP, Rubisco can also fix O_2, a reaction that competes with CO_2 fixation since both occurs in the same active site of the enzyme. However, while the fixation of CO_2 generates a three carbon intermediate (3-phosphoglicerate, PGA), which is further reduced to generate the end-product of photosynthesis, the triose phosphates (glyceraldehyde 3-phosphate, G3P), the fixation of O_2 produces both PGA and another metabolite, 2-phosphoglicolate (2-PG). Since 2-PG does not enter the PCR cycle and has no direct metabolic use, it is converted back to RuBP through a series of energy-dependent reactions. These reactions take place in the

chloroplast, peroxysomes and finally mitochondria, where CO_2 and NH_4^+ are released as by-products. This biochemical pathway is known as photorespiratory carbon oxidation, or photorespiration. The oxygenation reaction of Rubisco therefore decreases the efficiency of carbon fixation in three ways, by (1) preventing CO_2 to be fixed, since O_2 and CO_2 compete for the same active site in the enzyme, (2) consuming energy, and (3) releasing previously fixed CO_2 (Taiz and Zeiger 2010).

The overall impact of photorespiration on the net carbon assimilation increases with the relative O_2:CO_2 concentration, which is directly influenced by temperature (Bauwe et al. 2010). As temperature increases, CO_2 solubility decreases faster than O_2 solubility. Also, the CO_2:O_2 specificity of Rubisco decreases with increasing temperatures. In addition, stomatal closure, which can be triggered by factors such as aridity or high salinity, reduces CO_2 concentration in the leaf, thereby also influencing photorespiration rates. Photorespiration therefore imposes a stronger selective pressure in warm and dry habitats, where growth rates may be reduced as a consequence of higher rates of O_2 fixation by Rubisco. CO_2-concentrating mechanisms (CCMs) represent an evolutionary "work around" for the dual affinity of Rubisco for O_2 and CO_2 (Keeley and Rundel 2003; Edwards and Ogburn 2012). Their function increases the concentration of CO_2 relative to O_2 at the site of Rubisco, hence almost completely suppressing photorespiration and its associated costs. For this reason, CCMs confer an advantage in all habitats that promote high levels of photorespiration. Two major CCMs evolved in land plants over the ancestral photosynthetic pathway (i.e. C_3 photosynthesis), namely Crassulacean Acid Metabolism (CAM) and C_4 photosynthesis. These mechanisms are biochemically similar and rely on the separation of the initial fixation of atmospheric CO_2 and its ultimate fixation by Rubisco in the PCR cycle. This separation occurs temporally in CAM plants (with diurnal and nocturnal reactions), and spatially in C_4 plants (usually with two-cell reactions). A brief review of the C_4 photosynthetic pathway, which is the study system of this dissertation, is presented below.

The C_4 biochemical pathway was first described by Hatch and Slack (1966). However, various observations pointing out to an unusual carbon fixation mechanism in some tropical grasses preceded the formal description of the C_4 pathway (Furbank 2016). Hatch and Slack confirmed that malate and aspartate, and not PGA, were the first labelled products of $^{14}CO_2$ fixation in sugarcane leaves (Kortschak et al. 1965), and proposed a pathway which would eventually be known as C_4 photosynthesis.

The C_4 biochemistry relies on coupled metabolic reactions taking place in two cell types, the mesophyll (M) and bundle sheath (BS) cells (Fig. 1.5). The CO_2 that enters M cells is first equilibrated to bicarbonate (HCO_3^-) by carbonic anhydrase (CA) in the cytosol, and bicarbonate is incorporated into phosphoenolpyruvate (PEP) to generate a four-carbon compound, oxalacetate (OAA), in a reaction catalyzed by PEP carboxylase (PEPC). These two reactions are common to all C_4 plants (Furbank 2011). The fate of OAA and the pathway that releases CO_2 in the BS cells and regenerates PEP varies across taxa (Kanai and Edwards 1999). Generically, this OAA is converted into malate or pyruvate, which diffuses to BS cells, where it is decarboxylated by one or more of three possible decarboxylating enzymes: NADP-malic enzyme (NADP-ME), NAD-malic enzyme (NAD-ME) and PEP carboxykinase (PCK). The CO_2 that is released then feeds the PCR pathway, which in C_4 plants is confined to the BS cells, while a parallel pathway regenerates PEP in the M cells. Since BS cells have low diffusivity for CO_2 and are less exposed to the atmosphere than M cells, the C_4 cycle acts as a CO_2 pump, increasing the concentration of CO_2 relative to O_2, therefore reducing to insignificant levels the O_2 fixation by Rubisco, and the subsequent photorespiratory pathway (von Caemmerer and Furbank 2003).

The C_4 pathway is usually associated with a typical leaf anatomical arrangement, the so-called Kranz anatomy (Hattersley 1984; Dengler and Nelson 1999; Lundgren et al. 2014). This arrangement was firstly described long before the discovery of C_4 photosynthesis (Haberlandt 1884), and is generically characterized by short interveinal distances and large BS cells. BS and M cells of C_4 plants have major differences regarding their biochemical composition (Kanai and Edwards 1999), and position and number of organelles (Sage et al. 2014; Stata et al. 2014). Also, these cells are interconnected with higher densities of plasmodesmata in C_4 plants (Botha 1992; Danila et al. 2016, 2018), which facilitate intercellular metabolite diffusion. Although this constitutes the basal plan of C_4 anatomy, there is considerable variation in leaf anatomical traits across C_4 species (Hattersley 1984; Kellogg 1999; Soros and Dengler 2001; Kadereit et al. 2003; Lundgren et al. 2014). This includes a wide range of BS to M area, distance between veins, different numbers of bundle sheath cell layers, presence of additional photosynthetic cell types (Tateoka 1958; Renvoize 1986), and in the most exceptional case, species with a C_4 cycle that is performed among distinct compartments within the same cell (Voznesenskaya et al. 2001). The involvement of multiple anatomical and biochemical components make C_4 phenotype a typical complex trait,

which is moreover highly convergent and therefore constitutes an excellent study system to understand the genomic origins of complex traits.

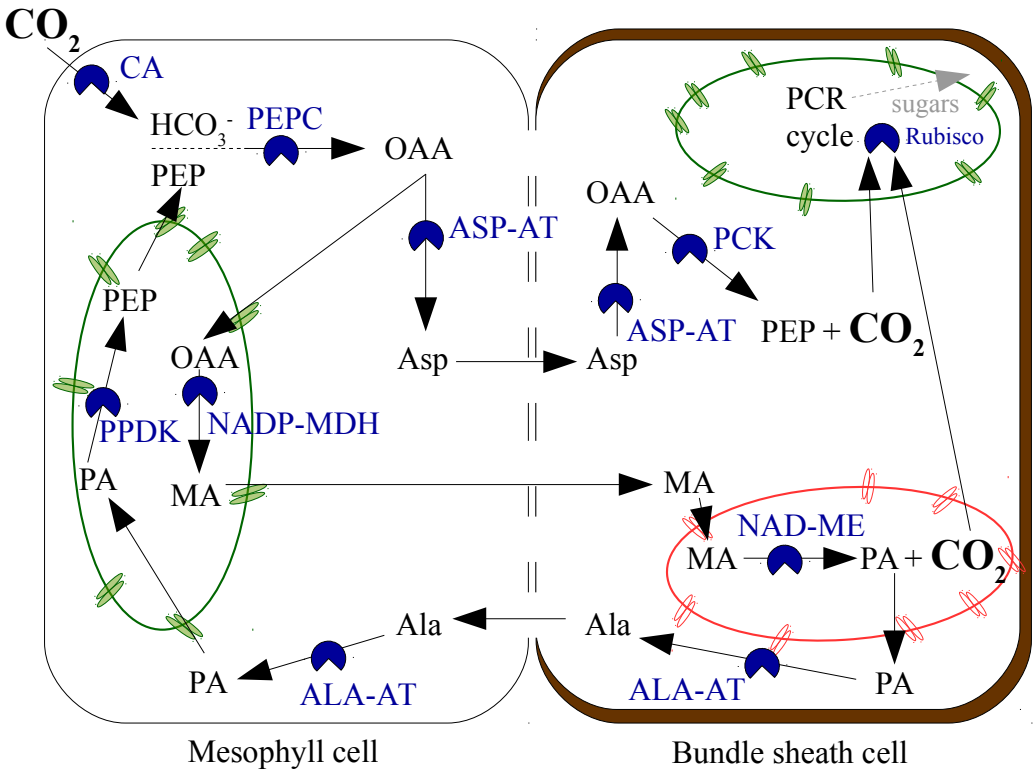


Fig. 1.5. Simplified representation of the C_4 biochemical cycle (NAD-ME subtype). Reactions take place in two cell types interconnected by plasmodesmata. Enzymes are represented as blue shapes, with their respective abbreviation: CA = carbonic anhydrase; ASP-AT = aspartate aminotransferase; ALA-AT = alanine aminotransferase; NAD-ME = NAD-dependent malic enzyme; NADP-MDH = NADP-dependent malate dehydrogenase; PCK = phosphoenolpyruvate carboxykinase; PEPC = phosphoenolpyruvate carboxylase; PPDK = phosphoenolpyruvate, pyruvate dikinase; Rubisco = ribulose 1,5 bisphosphate carboxylase/oxygenase. Metabolites: Asp = aspartate; HCO_3^- = bicarbonate; MA = malate; OAA = oxaloacetate; PA = pyruvate; PEP = phosphoenolpyruvate. PCR cycle is the photosynthetic carbon reduction (Calvin-Benson) cycle.

1.6. The evolutionary origins of C_4 photosynthesis

The C_4 metabolism is present in ~ 7,500 species in 19 families, accounting for ca. 3% of angiosperm diversity (Sage et al. 1999; Sage et al. 2011). Despite their relative low diversity, C_4 plants account for a quarter of the terrestrial gross primary production (Still et al. 2003). This is because large areas of the world are dominated by C_4 species, such as tropical and subtropical grasslands (Hartley 1958a; Hartley 1958b; Hartley and Slater 1960). C_4 has evolved independently at least 62 times (Sage et al. 2011), which makes it one of the most remarkable instances of convergent evolution in eukaryotes (Stern 2013). The first C_4 lineages evolved around 30 million years ago (Ma)

in the Oligocene, which coincides with a drop in the global atmospheric CO_2 concentration (Ehleringer et al. 1991; Christin et al. 2008; Vicentini et al. 2008). This drop is suggested to be a major selective pressure on the evolution of C_4 photosynthesis, since it provided the conditions for C_4 to be energetically advantageous over C_3 plants in habitats that promote photorespiration (Ehleringer and Bjorkman 1977). The dominance of C_4 plants, however, happened ~ 20-25 million years later, during the Late Miocene, which is characterized by the expansion of C_4 grasslands (Quade et al. 1989; Cerling 1999; Osborne 2008).

Despite their multiple convergent origins, C_4 species are not evenly distributed across the phylogeny of land plants. Most of the C_4 origins occurred in a few clades, such as grasses (> 22 origins; GPWG II 2012), sedges (> five origins; Besnard et al. 2009), and Caryophyllales (> 15 origins; Sage et al. 2011). Such clustered phylogenetic distribution points to the occurrence of preconditions for C_4 evolution, which might be present in only some phylogenetic groups (Sage 2001; Christin and Osborne 2014). In terms of leaf anatomy, for example, it was shown that grass lineages with a proportion of BS tissue in the leaf that is higher than 15% account for a significantly higher number of C_4 origins (Christin et al. 2013). Recent evidence from studies on molecular evolution suggests that some genetic preconditions also exist. Independent C_4 origins co-opted the same genes from multiple copies present in the genomes of C_3 ancestors, which suggests that some members of gene families are more suitable than others for the C_4 function (Christin et al. 2007; Christin et al. 2009; Brown et al. 2011; Williams et al. 2012; Christin et al. 2013; Moreno-Villena et al. 2018).

The multiple origins of C_4 photosynthesis might suggest that the C_4 pathway is easy to evolve, if, for example, its emergence were mediated by a master key regulator. However, individual C_4-related genes are regulated at multiple levels, and these regulatory mechanisms can vary across C_4 lineages (Hibberd and Covshoff 2010; Langdale 2011; Wang et al. 2011). Nonetheless, recent studies reported that independent C_4 lineages express common regulatory elements (Aubry et al. 2014; John et al. 2014; Reyna-Llorens 2018). These findings suggest that either these regulatory sequences were already present in the common C_3 ancestor or they evolved independently through parallel, progressive genetic modifications. Gene regulation of leaf development has been intensively studied, particularly due to efforts to engineer C_4 photosynthesis into C_3 plants (von Caemmerer et al. 2012). The GOLDEN2 gene has an important role in the formation of bundle-sheath cells in maize (Hall et al. 1998), and recently the

constitutive expression of the transcription factors GOLDEN2-like in the C_3 rice induced a proto-Kranz leaf anatomy (Wang et al. 2017). These recent findings suggest that one or a few loci might control the development of C_4 leaf anatomy and tissue-specific expression patterns. However, the other components of C_4 photosynthesis are controlled by different genetic modifications, so that the C_4 trait remains a complex phenotype that relies on the coordinated action of multiple anatomical and biochemical components.

The order in which the components of C_4 photosynthesis evolved is still debated. The existence of so-called C_3-C_4 intermediate species provides clues on how this transition might have happened (Monson and Moore 1989; Sage 2004, 2012). The C_3-C_4 intermediate photosynthetic metabolism is characterized by a photorespiratory cycle that occurs in two separate cells (Ku et al. 1983; Monson et al. 1984). In these plants, a large proportion of the CO_2 that is released via photorespiration is refixed by Rubisco. This metabolism, known as C_2, or photorespiratory CO_2-recycling, is present in lineages that have in some cases closely related members that are C_4, which also suggests that this phenotype might be an evolutionary link between the C_3 and C_4 states. Biochemical modeling suggested that a transition between the C_2 and C_4 states would result from the need to rebalance nitrogen metabolism between the different types of cells in a process that involves enzymes that are part of the C_4 cycle (Mallmann et al. 2014). It is furthermore predicted that any increase of the strength of a rudimentary C_4 cycle will result in biomass gains, and therefore a fitness advantage (Heckmann et al. 2013). This model was however based on a single C_4 origin, in the eudicot *Flaveria* (Asteraceae), and evidence from comparative analyses are lacking. The grass family accounts for a high number of C_4 origins, and is the most ecologically successful and economically relevant group of C_4 plants. Grasses therefore represent an outstanding model group for understanding C_4 evolution.

1.7. Grasses as a model group

The grass family (Poaceae) comprises more than 11,000 species with a cosmopolitan distribution (Clayton and Renvoize 1986; Watson and Dallwitz 1992 onwards; Kellogg 2015). The group encompasses a large number of domesticated species, including the three major food crops in the world - rice, wheat and maize -, some biofuel crops (e.g. sugar cane and *Miscanthus* spp.) and various forage species used in pastures. Grasses

also dominate vast natural areas of tropical and temperate regions, including grasslands in African, the prairies of North America, the pampas of South America, and the bamboo forests in Asia (Kellogg 2015). Such astonishing ecological success and significance for human societies demonstrates the importance of this plant family.

Grasses are taxonomically divided into five phylogenetic major groups: the three species-poor, early-diverging subfamilies Anomochlooideae, Pharoideae and Puelioideae, and the two major clades (crown grasses) BOP and PACMAD (Fig. 1.6). The BOP and PACMAD clades are similarly sized and include most of the species diversity in the family. The BOP clade encompasses most of the cold-adapted species in the family in addition to bamboos and rice and its relatives (Hartley 1961; 1973; Kellogg 2015). The PACMAD diversity is mainly distributed across tropical and subtropical regions, but with several lineages that colonized colder regions secondarily (Hartley 1958; Hartley and Slater 1960; Kellogg 2015). The divergence between the two clades probably happened between 50 and 60 Ma, based on molecular dating (Vicentini et al. 2008; Bouchenak-Khelladi et al 2014; Jones et al. 2014; Christin et al. 2014). This date would however be pushed back to ~ 80 Ma if controversial phytolith calibration points were used for the estimation (Prasad et al. 2005; Christin et al. 2014; Kellogg 2015).

The ecological success of some dominant grasses in open and drier habitats is generally associated with the C_4 metabolism (Long 1999), which occurs in at least 4,500 grass species. All C_4 grasses belong to the PACMAD clade, and this is the most species-rich C_4 group of angiosperms, with also the largest number of independent origins of the trait (22-24 times; GPWG II 2012; Fig. 1.6). The oldest C_4 lineages evolved around 25-35 Ma in the Chloridoideae subfamily, with other origins spread throughout the last 25 million years (Christin et al. 2008). The recurrent origins of C_4 photosynthesis within grasses constitute an outstanding system for comparative studies addressing the ecological and physiological origins and consequences of this shift in photosynthetic type (Edwards et al. 2008, 2010; Osborne and Freckleton 2009; Spriggs et al. 2014; Atkinson et al. 2016; Watcharamongkol et al. 2018). In addition, these recurrent origins coupled with extensive genomic resources for the family made it perfectly suited to address the repeatability of adaptive changes at the molecular level (Christin et al. 2007, 2009). However, previous studies of C_4 origins in grasses mainly relied on single gene/enzyme analyses. In addition, the temporal scale of C_4 evolution in this group can blur some of the early events.

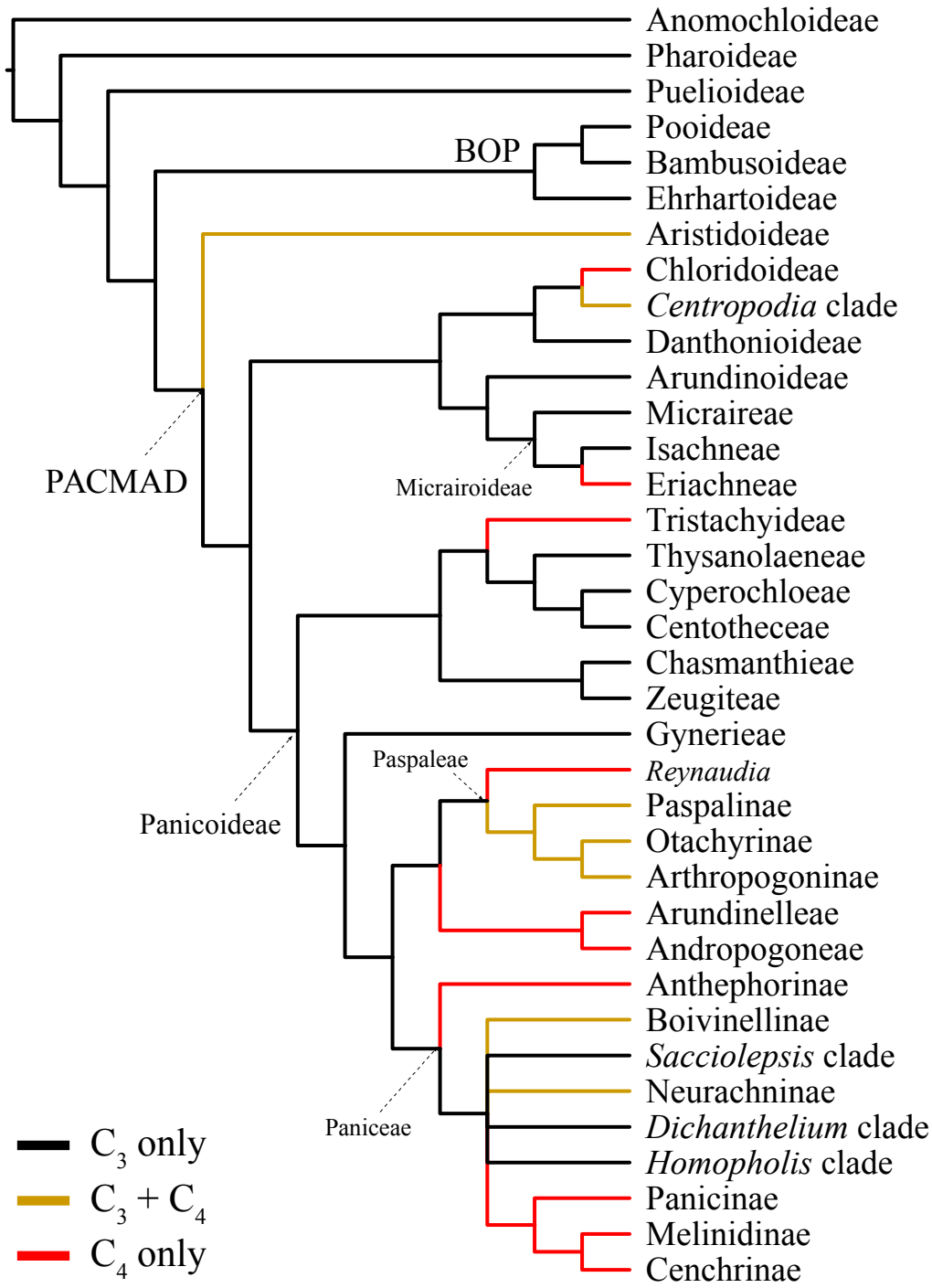


Fig. 1.6. Phylogenetic relationships within the grass family (Poaceae) based on plastid markers. Tree redrawn from GPWG II (2012).

Besides the old, large C_4 clades, the grass family includes recent C_4 lineages, some of which contain C_3-C_4 intermediates. This is the case of *Steinchisma* and *Neurachne*, for the latter of which includes C_3, C_3-C_4 and C_4 species (Duvall et al. 2003; Christin et al. 2012b). However, the most promising grass to address the early events during C_4 evolution might be *Alloteropsis semialata*. Although most species of the

genus *Alloteropsis* (Boivinellinae, Panicoideae) are C_4, the species *A. semialata* includes C_4, C_3 and C_3-C_4 intermediate populations (Ellis 1974; Lundgren et al. 2015; Lundgren et al. 2016). *A. semialata* corresponds to one of the most recent divergences of photosynthetic types in grasses (Christin et al. 2012a; GPWG II 2012; Dunning et al. 2017), and provides, therefore, a remarkable model to investigate the evolution of the complex C_4 trait. C_3 and C_4 individuals have been compared in the past to address the ecophysiological consequences of C_4 photosynthesis (e.g. Ripley et al. 2007; Ibrahim et al. 2008; Osborne et al. 2008; Ripley et al. 2008), and recent intraspecific investigations are revealing the history of transitions within *A. semialata* and the complex it forms with its congeners (Lundgren et al. 2015; Dunning et al. 2017). However, before the present thesis, comparative genomic analyses had been conducted only with transcriptome data and with only one C_3 and one C_4 representative (Christin et al. 2013). This genus therefore represents an outstanding system to address the genomic origins of the C_4 complex trait, while comparison with other grass lineages can then determine the generality of the observed patterns.

1.8. Thesis Aims and Structure

The overarching aim of this dissertation is to identify genomic changes associated with the evolutionary origin of complex traits. This effort is conducted using C_4 photosynthesis as a study system. Whole genome sequencing data of selected grass lineages are analysed in a phylogenetic framework to address several inter-related questions, which together aim to depict the evolutionary processes that lead to the realization of complex physiological novelties.

In Chapter 2, I collaborate with colleagues from the University of Toulouse, in France, to investigate the tempo and the order of changes involved in the transition to a C_4 physiology in the Andropogoneae grasses. This species-rich lineage encompasses multiple economically and ecologically important species, and the evolution of C_4 is assumed to be a key innovation for the ecological success of the group. Although this group includes some of the most studied C_4 species (maize, sorghum and sugarcane), previous comparative studies missed the changes leading to the evolution of the C_4 trait because of the lack of a known closely related C_3 lineage. In this chapter, I develop novel phylogenomic approaches to verify whether a rare C_3 lineage from India, sampled by my collaborators, is sister to Andropogoneae. Sequence evolution of C_4 enzymes is then analysed to determine whether adaptive changes happened at the base of Andropogoneae, or were spread along the diversification of the group. Additionally, leaf anatomical data are compiled to assess diversity in the components of the C_4 trait within Andropogoneae.

The insights gained in Chapter 2 are limited to interspecies comparisons due to the ancient origin of C_4 photosynthesis in the Andropogoneae, which blurs the microevolutionary changes associated with the trait. In Chapter 3, I overcome this limitation by performing an intraspecific study, this time using an emergent C_4 model, the grass *Alloteropsis semialata*, which includes C_4 and non-C_4 populations, and in which C_4 evolved more recently, at ca. 3 Ma. To investigate the molecular mechanisms underlying the transition to a C_4 physiology, I test a novel hypothesis for the role of gene duplication during C_4 evolution. I develop a novel approach to estimate gene copy numbers from low-coverage genome datasets, and correlate it with published

transcriptome data to test whether gene duplication can provide a mechanism to rapidly increase the expression of C_4 genes, therefore facilitating the evolution of a C_4 physiology. I further investigate whether the acquisition of genes better adapted for the C_4 function (with the example of laterally acquired genes) supersedes the dosage effect of gene duplication.

The results from Chapter 3 highlight the importance of some laterally-acquired genes for the emergence of C_4 photosynthesis in *A. semialata*, and also the convenient study system that they represent. In Chapter 4, I capitalize on this system to investigate the order of acquisition of C_4 components. Using low-coverage genome data, I assemble alleles of laterally-acquired genes encoding two key C_4 enzymes to reconstruct the history of gene flow among populations of *A. semialata.* This work is performed in a population genomic framework developed by Dr. Jill Olofsson, a postdoctoral research associate from our research group with whom I collaborated for this chapter. With a detailed picture of the phylogeographical distribution of C_4 laterally-acquired genes, we then test whether components for the C_4 pathway can be acquired independently and later combined during secondary contacts.

Chapter 4 sheds new light on the history of laterally-acquired genes within *A. semialata*, but these investigations are conducted on a very large scale with relatively few samples. In Chapter 5, I sequence and assemble the genome of an *A. semialata* individual to track the post-acquisition spread of one of these genes in detail. New accessions from candidate donors are examined to narrow the timing and geographical region where the event of lateral transfer has occurred, and the fragment containing the laterally acquired C_4 gene in *A. semialata* is characterized, along with the circumstances of its transfer and spread to different populations. With the example of lateral gene transfers, this chapter highlights the role of intraspecific gene movements for building the trait diversity and complexity accumulated throughout its evolutionary history.

Overall, each of the four studies contributes to fill in the knowledge gaps that produce the apparent conundrum of evolving complex physiological traits. The findings presented here suggest that most of the complexity observed in the C_4 trait across species is the result of lineage-specific increments on top of an initial C_4 physiology, rather than the necessary steps of a single route that culminates in an optimal, fully C_4

physiology. This thesis also develops new methodological approaches to extract meaningful and robust biological information from the often neglected nuclear genome sequences of low-coverage sequencing datasets. In this sense, this work provides methods to gain valuable insights into a complex biological problem at relatively low costs and in ways that are not restricted to model species.

Chapter 2.

Continued adaptation of C_4 photosynthesis after an initial burst of changes in the Andropogoneae grasses

Chapter 2. Continued adaptation of C_4 photosynthesis after an initial burst of changes in the Andropogoneae grasses

Matheus E. Bianconi[1*], Jan Hackel[2*], Maria S. Vorontsova[3], Adriana Alberti[4], Watchara Arthan[5], Sean V. Burke[6], Melvin R. Duvall[6], Elizabeth A. Kellogg[7], Sébastien Lavergne[8], Michael R. McKain[9], Alexandre Meunier[2], Colin P. Osborne[1], Paweena Traiperm[5], Pascal-Antoine Christin[1], Guillaume Besnard[2]

[*] These authors contributed equally to this work

[1] Department of Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, UK.

[2] Laboratoire Evolution & Diversité Biologique (EDB, UMR 5174), CNRS/IRD/Université Toulouse III, 118 route de Narbonne, 31062 Toulouse, France.

[3] Comparative Plant and Fungal Biology, Royal Botanic Gardens, Kew, Richmond, Surrey TW93AB, UK.

[4] CEA - Institut de Biologie Francois-Jacob, Genoscope, 2 Rue Gaston Cremieux 91057 Evry Cedex, France

[5] Department of Plant Science, Faculty of Science, Mahidol University, King Rama VI Road, Bangkok 10400, Thailand.

[6] Plant Molecular and Bioinformatics Center and Department of Biological Sciences, Northern Illinois University, 1425 W. Lincoln Hwy, DeKalb, IL 60115-2861, USA.

[7] Donald Danforth Plant Science Center, 975 North Warson Road, St. Louis, Missouri 63132, USA.

[8] Laboratoire d'Ecologie Alpine, CNRS – Université Grenoble Alpes, UMR 5553, Grenoble, France.

[9] Department of Biological Sciences, The University of Alabama, 500 Hackberry Lane, Tuscaloosa, AL 35487, USA.

Personal contribution: I performed the nuclear genome analyses and generated the phenotypic data. I co-wrote the manuscript with Dr. Jan Hackel, who performed the plastome and biogeographic analyses. Drs. Maria S. Vorontsova and Elizabeth A. Kellogg did the taxonomic treatment. All co-authors commented on the text.

2.1. Abstract

C_4 photosynthesis is a complex trait that sustains fast growth and high productivity in tropical and subtropical conditions and evolved repeatedly in flowering plants. One of the major C_4 lineages is Andropogoneae, a group of ~ 1,200 grass species that includes some of the world's most important crops and species dominating tropical and some temperate grasslands. Previous efforts to understand C_4 evolution in the group have compared a few model C_4 species to distantly related C_3 species, so that changes directly responsible for the transition to C_4 could not be distinguished from those that preceded or followed it. In this study, we developed a new approach to analyse the genomes of 98 grass species that captures the earliest diversification within Andropogoneae as well as their C_3 relatives. Phylogenomics combined with molecular dating and positive selection tests show that many changes linked to the evolution of C_4 photosynthesis happened in the Early Miocene, between 21 and 18 Ma, preceding the initial diversification of Andropogoneae. This initial burst of changes was followed by an extended period of modifications to leaf anatomy and biochemistry during the diversification of Andropogoneae, so that a single C_4 origin gave birth to a diversity of C_4 phenotypes during 18 million years of speciation events and migration across geographic and ecological space. Our innovative approach and sampling of the diversity in the group reveals that one key transition can lead to a plethora of phenotypes following sustained adaptation of the ancestral state.

Keywords: adaptive evolution, Andropogoneae, C_4 photosynthesis, complex traits, herbarium genomics, Jansenelleae, Poaceae, phylogenomics.

2.2. Introduction

One of the major goals of evolutionary biology is to understand the origins of key innovations underlying the ecological success of some groups. This requires the study of adaptive traits in a phylogenetic context, using comparisons of species differing in character states (e.g. Bond et al. 2014; Rainford et al. 2014; McGee et al. 2015; Sánchez-García and Matheny 2017). Because species differ in numerous ways, such comparisons must capture the diversity that emerged after the transition in addition to the diversity that preceded it, to precisely identify the properties directly involved in the origin of the trait of interest (e.g. Sprent 2007; Yukawa 2009; Endress 2011; Puttick 2014; Marek and Moore 2015; Clark et al. 2018). Among flowering plants, C_4 photosynthesis represents an adaptive novelty with significant ecological consequences (Sage 2004; Edwards et al. 2010; Christin and Osborne 2014).

C_4 photosynthesis results from multiple anatomical and biochemical modifications of the ancestral C_3 photosynthetic metabolism, which include (1) the confinement of the primary enzyme of the photosynthetic carbon reduction pathway, Rubisco, to a compartment isolated from the atmosphere, and (2) increased and cell-specific activity of several enzymes that increase CO_2 concentration at the site of Rubisco (Hatch 1987; von Caemmerer and Furbank 2003; Sage 2004). The concentration of CO_2 around Rubisco boosts photosynthetic efficiency, and therefore growth, particularly in high-light, warm and dry conditions (Long 1999; Atkinson et al. 2016).

Although the C_4 trait requires the modification of multiple components, it has evolved at least 62 times independently during the diversification of angiosperms (Sage et al. 2011). The grass family (Poaceae) encompasses almost half of the C_4 origins, including some with particular ecological and economic relevance, such as the Andropogoneae (Sage 2017). The roughly 1,200 species of this tribe are all C_4, making it the second most speciose C_4 lineage (Sage et al. 2011). Andropogoneae include some of the world's most important cereal and fuel crop plants, such as maize, sorghum, sugarcane and *Miscanthus* spp. (e.g. silver grass), as well as numerous dominant species of tropical and some temperate grasslands (Hartley 1958; Bond et al. 2008; Edwards et al. 2010; Kellogg 2015). Besides generating some of the most productive plants in the world, their C_4 trait also increased the diversification of Andropogoneae, which in turn have shaped ecosystems around the world (Osborne 2008; Edwards et al. 2010;

Forrestel et al. 2014; Spriggs et al. 2014; Sage and Stata 2015).

Due to their economic and ecological importance, Andropogoneae have been included in most studies addressing the evolutionary origins of C_4 photosynthesis in grasses. In particular, efforts to determine the genomic changes involved in the transition to C_4 photosynthesis have focused on comparisons between the two C_4 model Andropogoneae species – maize and sorghum – and distantly related C_3 model grasses (e.g. rice and *Dichanthelium*; Paterson et al. 2009; Wang et al. 2009; Emms et al. 2016; Studer et al. 2016; Huang et al. 2017). Such a narrow taxon sampling neither covers the diversity of anatomical and biochemical components observed among C_3 grasses (e.g. Hattersley 1984; Christin et al. 2013; Lundgren et al. 2014) nor among C_4 grasses within the Andropogoneae themselves (Renvoize 1982a; Ueno 1995; Sinha and Kellogg 1996). However, sampling this diversity is crucial for determining those modifications involved in the early origin of the C_4 pathway in the group as opposed to its subsequent diversification (Christin et al. 2010; Dunning et al. 2017a). Furthermore, as with other key innovations, it remains unclear whether C_4 evolution was concentrated during a short episode of transition or occurred over a prolonged period of improvement during the diversification of C_4 clades. While determining the extent of changes required for the initial transition to C_4 would help understanding how such a complex trait evolved so many times independently, it might also inform current efforts to engineer C_4 photosynthesis in rice (von Caemmerer et al. 2012).

The first divergence within Andropogoneae *sensu* Kellogg (2015) separates the subtribe Arundinellinae from Andropogoneae *s.s.* (tribes Arundinelleae and Andropogoneae, respectively in Soreng et al. 2017), the latter of which includes the model species maize and sorghum. Until recently, no closely related C_3 sister lineage was known for Andropogoneae (GPWG II 2012), but such a position has been suggested for the C_3 genera *Jansenella* and *Chandrasekharania* based on individual chloroplast or nuclear markers (Besnard et al. 2018; Hackel et al. 2018). Genomic resources are extremely sparse for species from these genera and Arundinellinae, but low-coverage genome scans have recently provided insights into the evolution of the nuclear genome in other non-model grasses (Besnard et al. 2014, 2018; Chapter 3; Chapter 4). Capitalizing on the availability of such genomic datasets as a side-product of plastome sequencing (e.g. Washburn et al. 2015; Burke et al. 2016; Arthan et al. 2017; Piot et al. 2018), we are now able to phylogenetically track the modifications

underlying one of the major innovations of flowering plants.

In this study, we combine novel approaches to analyse genome scans for a large number of grasses covering the diversity of C_3 relatives of Andropogoneae as well as the earliest diversification within the group. First, we generate the first genome-wide nuclear phylogeny of grasses to confirm relationships among Andropogoneae and with their C_3 relatives, and compare it to a new plastome phylogeny for the group. Next, using molecular dating, we estimate the age of C_4 photosynthesis in Andropogoneae. Anatomical traits are then mapped onto the time-calibrated phylogeny to infer leaf structural transitions in the group. Finally, we use positive-selection analyses to detect episodes of adaptive evolution in key C_4 enzymes, testing the alternative hypotheses that adaptive changes (i) occurred in a C_3 context and therefore predated the origin of Andropogoneae, (ii) occurred at the base of the clade, during a short period of time, representing the major transition from C_3 to C_4 photosynthesis, or (iii) were sustained throughout the history of the group, representing a prolonged period of gradual innovation within the monophyletic C_4 Andropogoneae. Overall, our study presents new approaches to dissecting a complex adaptive trait and analysing its components in isolation, to infer the tempo of key phenotypic transitions in a large group of ecological importance.

2.3. Material and Methods

2.3.1. Species sampling, sequencing and distribution

A dataset of low-coverage genome sequences was assembled that covers the main lineages of Andropogoneae including the subtribe Arundinellinae and the Andropogoneae *s.s.* (*sensu* Kellogg 2015), which represents the earliest known split within this C_4 group (GPWG II 2012); their putative closest C_3 relatives; a variety of other C_3 and C_4 Panicoideae; and representatives of the other grass subfamilies (Table 2.1). In total, genomic data for 90 grass species were retrieved from previous studies, and similar data for eight species were generated here (Table 2.1). For the latter, low-coverage sequencing was performed using Illumina technology. Genomic DNA (gDNA) was isolated from ca. 5–10 mg of leaf material using the BioSprint 15 DNA Plant Kit (Qiagen). Libraries were prepared from 200–500 ng of gDNA using the Illumina TruSeq Nano DNA Sample Prep Kit. Fragments were either sonicated or size-selected (50–300 bp of insert size) and enriched with 12 PCR cycles using the proofreading polymerase

supplied with the Illumina kit (the latter specifically for libraries prepared from herbarium material; Table 2.1). Paired-end sequencing of the genomic libraries was performed either on a HiSeq 2000, 2500 or 3000 at the Genopole Toulouse or at the Genoscope Évry platforms in France.

The geographic distribution of Andropogoneae and their C_3 relatives was assessed with species-level distribution data for the World Geographical Scheme for Recording Plant Distributions (TDWG) level-3 botanical regions (corresponding largely to countries or states), retrieved from the World Catalogue of Selected Plant Families (Clayton et al. 2016). Numbers of species and endemics per botanical region were plotted with the package *rgdal* (Bivand et al. 2017) in R 3.4.3 (R Core Team 2017), using shapefiles provided by the Royal Botanic Gardens, Kew (https://www.kew.org/gis/tdwg/index.html, accessed on 10 August 2017).

2.3.2. Plastome analysis

Full plastome sequences were either retrieved from NCBI or assembled in this study using the genomic datasets (Table 2.1). For those assembled here, either a *de novo* strategy was used with the software OrgAsm v.1.0 (http://pythonhosted.org/ORG.asm), or a consensus sequence was called by mapping reads to a closely related reference plastome (where available) using Geneious v.9.1.8 (Kearse et al. 2012) and extending/reducing indels by repeated mapping to contigs where necessary. Potential errors in the *de novo* assembly were corrected by mapping the genomic reads to the assembled sequence in Geneious, from which a new consensus sequence was called using the highest-quality base criterion. Mean estimated sequencing depth ranged from 90 to 4602 reads per site.

The 98 plastome sequences were aligned with MAFFT v.7.13 (Katoh and Standley 2013), after excluding the second inverted repeat region to avoid representing the same sequence twice. Maximum likelihood (ML) plastome trees were inferred from the 147,601 bp alignment using RAxML v.8.2.4 (Stamatakis 2014) with a GTR+CAT substitution model and assessing node support with 1000 rapid bootstrap pseudoreplicates. Trees were rooted using the BOP clade (Bambusoideae, Oryzoideae and Pooideae) as outgroup (GPWG II 2012).

Table 2.1. Genomic data information.

Species	Voucher/isolate/	Subfamily/Tribe	Source study	SRA	Plastome accession
Alloteropsis angusta	Pauwels 1182 (BR)	Paniceae	Chapter 4	SRP082653	KX752090.1
Alloteropsis cimicina	Hall 20 (K)	Paniceae	Lundgren et al. 2015	SRP082653	NC_027952.1
Alloteropsis semialata	AusTRCF 322458 0167	Paniceae	Lundgren et al. 2015	SRP082653	KT281145.1
Amphicarpum muhlenbergianum	Clark et al. 1695 (ISC)	Paniceae	Burke et al. 2016	-	NC_030619.1
Andropogon burmanicus	Arthan 071	Andropogoneae	Arthan et al. 2018	-	KY596164.1
Aristida rufescens	MSV330	Aristidoideae	Piot et al. 2017	-	MF563384.1
Arthraxon microphyllus	Traiperm 537	Andropogoneae	Arthan et al. 2018	-	KY596183.1
Arthraxon prionodes	PI<ITA>:659331	Andropogoneae	Burke et al. 2016	-	NC_030613.1
Arundinella deppeana	Clark et al. 1680 (XAL)	Andropogoneae	Burke et al. 2016	-	NC_030620.1
Arundinella hirta	USDA PI 246756	Andropogoneae	Washburn et al. 2015	SRR2163563	not submitted
Arundinella hookeri	Kew #0050290	Andropogoneae	Washburn et al. 2015	SRR2163560	not submitted
Arundinella nepalensis	MSV608	Andropogoneae	This study	-	not submitted
Axonopus fissifolius	Clark et al. 1703 (ISC)	Paspaleae	Burke et al. 2016	-	NC_030501.1
Bothriochloa alta	DEK:Duvall s.n.	Andropogoneae	Burke et al. 2016	-	NC_030621.1
Brachiaria fragrans	JMB19160	Paniceae	Silva et al. 2017	-	KX663837.1
Brachypodium distachyon	SAMN05519009	Pooideae	NCBI SRA Archive	SRR4029428	NC_011032.1
Capillipedium venustum	PI<ITA>:11713	Andropogoneae	Burke et al. 2016	-	NC_030622.1
Chandrasekharania keralensis	VSR54064	Jansenelleae	Besnard et al. 2018	-	not submitted
Chasechloa egregia	LHB s.n.	Paniceae	Silva et al. 2017	-	KX663836.1
Chasechloa madagascariensis	HPB11217	Paniceae	Silva et al. 2017	-	KX663838.1
Chasmanthium sessiliflorum	ISC(USA-IA):Sanchez-Ken	Chasmanthieae	Burke et al. 2016	-	KU291494.1
Chrysopogon zizanioides	Kellogg Vet-MRL-001	Andropogoneae	Arthan et al. 2018	-	KY596158.1
Coix lacryma-jobi	Arthan 072	Andropogoneae	Arthan et al. 2018	-	KY596160.1

Coleachne africana	RCH09	Micrairoideae	Piot et al. 2017	-	MF563382.1
Cymbopogon citratus	09Cc	Andropogoneae	Dunning et al. In prep	-	not submitted
Danthoniopsis dinteri	SRR2163566	Tristachyideae	Burke et al. 2016	-	NC_030502.1
Danthoniopsis stocksii	RHR54967	Tristachyideae	This study	-	not submitted
Dichanthelium acuminatum	Saarela 666 (CAN)	Paniceae	Burke et al. 2016	-	NC_030623.1
Dichanthium aristatum	08Da	Andropogoneae	Dunning et al. In prep	-	not submitted
Digitaria glauca	MSV950	Paniceae	This study	-	not submitted
Diheteropogon ampelectens	PI<ITA>:12585	Andropogoneae	Burke et al. 2016	-	KU291497.1
Dimeria ornithopoda	Traiperm 575	Andropogoneae	Arthan et al. 2018	-	KY596130.1
Echinochloa stagnina	RCH49	Paniceae	Piot et al. 2017	-	MF563380.1
Eremochloa ciliaris	Traiperm 524	Andropogoneae	Arthan et al. 2018	-	KY596146.1
Eriochloa meyeriana	Duvall s.n. (DEK)	Paniceae	Burke et al. 2016	-	NC_030624.1
Eriochrysis cayennensis	Welker 365	Andropogoneae	Arthan et al. 2018	-	NC_029882.1
Eulalia aurea	PI<ITA>:12153	Andropogoneae	Burke et al. 2016	-	NC_030503.1
Eulalia siamensis	Traiperm 557	Andropogoneae	Arthan et al. 2018	-	KY596149.1
Garnotia stricta var. longiseta	RS1386	Andropogoneae	Besnard et al. 2018	-	not submitted
Garnotia tenella	Traiperm 552	Andropogoneae	Arthan et al. 2018	-	KY596184.1
Garnotia thailandica	Traiperm 535	Andropogoneae	Arthan et al. 2018	-	KY596171.1
Glyphochloa forficulata	HFP896	Andropogoneae	This study	-	not submitted
Gynerium sagittatum	P6-301b	Gynerieae	This study	-	not submitted
Hildaea pallens	GB06-2014	Paspaleae	Piot et al. 2017	-	MF563377.1
Homolepis aturensis	GB06-2012	Paspaleae	Piot et al. 2017	-	MF563378.1
Hyparrhenia subplumosa	PI<ITA>:12665	Andropogoneae	Burke et al. 2016	-	NC_030625.1
Imperata cylindrica	DEK:Burke 21	Andropogoneae	Burke et al. 2016	-	NC_030487.1
Ischaemum afrum	PI<ITA>:364924	Andropogoneae	Burke et al. 2016	-	NC_030488.1

Iseilema macratherum	PI<ITA>:257760	Andropogoneae	Burke et al. 2016	-	NC_030611.1
Iseilema membranaceum	07Im	Andropogoneae	Dunning et al. In prep	-	not submitted
Jansenella griffithiana	CG209	Jansenelleae	Besnard et al. 2018	-	not submitted
Jansenella neglecta	SRY201	Jansenelleae	This study	-	submit
Kerriochloa siamensis	Traiperm 580	Andropogoneae	Arthan et al. 2018	-	KY596120.1
Lasiacis nigra	GB02-2014	Paniceae	Piot et al. 2017	-	MF563376.1
Lasiorhachis hildebrandtii	LRK2008	Andropogoneae	Piot et al. 2017	-	MF563371.1
Lasiurus scindicus	AN-DJI/78-36	Andropogoneae	Besnard et al. 2018	-	not submitted
Lecomtella madagascariensis	MSV603	Lecomtelleae	Besnard et al. 2013	-	HF543599.2
Loudetia simplex	MSV1049	Tristachyideae	Piot et al. 2017	-	MF563366.1
Loudetiopsis kerstingii	PI<ITA>:12679	Tristachyideae	Burke et al. 2016	-	NC_030612.1
Megathyrsus maximus	PI 12181	Paniceae	Burke et al. 2016	-	NC_030489.1
Melinis minutiflora	MSV609	Paniceae	This study	-	not submitted
Merxmuellera tsaratananensis	MSV486	Danthonioideae	Piot et al. 2017	-	MF563375.1
Miscanthus sinensis	SAMN01163223	Andropogoneae	NCBI SRA Archive	SRP015486	NC_028721.1
Mnesithea helferi	Traiperm 574	Andropogoneae	Arthan et al. 2018	-	KY596162.1
Oncorachis ramosa	Zuloaga 6960	Paspaleae	Burke et al. 2016	-	NC_030490.1
Oplismenus hirtellus	Clark & Lewis 1644 (ISC)	Paniceae	Burke et al. 2016	-	NC_030491.1
Oryza sativa	SAMD00045947	Oryzoideae	NCBI SRA Archive	DRR054198	X15901.1
Otachyrium versicolor	Zuloaga 7027	Paspaleae	Burke et al. 2016	-	NC_030492.1
Panicum capillare	Saarela 769 (CAN)	Paniceae	Burke et al. 2016	-	NC_030493.1
Panicum lycopodioides	GB04-2013	Paniceae	Piot et al. 2017	-	MF563374.1
Paspalidium geminatum	Giussani 313 (SI)	Paniceae	Burke et al. 2016	-	NC_030494.1
Paspalum paniculatum	MSV500	Paspaleae	Piot et al. 2017	-	MF563367.1
Paspalum vaginatum	SAMN05660222	Paniceae	NA	SRR4136360	not submitted
Phyllostachys edulis	SAMN03417478	Bambusoideae	NA/Zhang et al. 2015	SRR1916022	NC_015817.1

Plagiantha tenella	Zuloaga 6953	Paspaleae	Burke et al. 2016	-	NC_030497.1
Polytoca digitata	Arthan 054	Andropogoneae	Arthan et al. 2018	-	KY596178.1
Pseudolasiacis leptolomoides	MSV983	Paniceae	Piot et al. 2017	-	MF563372.1
Pseudosorghum fasciculare	Arthan 067	Andropogoneae	Arthan et al. 2018	-	KY596157.1
Reynaudia filiformis	E12208	Paspaleae	This study	-	not submitted
Rottboellia cochinchinensis	ISC(USA-IA)Clark et al. 1698	Andropogoneae	Burke et al. 2016	-	NC_030615.1
Sartidia dewinteri	SDW698	Aristidoideae	Besnard et al. 2014	-	KJ819550.1
Sartidia isaloensis	MSV1325	Aristidoideae	Piot et al. 2017	-	MF563370.1
Sartidia perrieri	HPB10751	Aristidoideae	Besnard et al. 2014	-	KJ819549.1
Setaria italica	SRA pooled samples	Paniceae	NCBI SRA Archive	-	KJ001642.1
Sorghastrum nutans	DEK:Wysocki s.n.	Andropogoneae	Burke et al. 2016	-	NC_030498.1
Sorghum bicolor	SAMN05519228	Andropogoneae	NCBI SRA Archive	SRR4028749	NC_008602.1
Sporobolus michauxianus	SAMN05920556	Chloridoideae	NA	SRR4434179	NC_029416.1
Steinchisma laxum	Zuloaga 7416	Paspaleae	Burke et al. 2016	-	NC_030499.1
Stipagrostis hirtigluma	MSV902	Aristidoideae	Piot et al. 2017	-	MF563365.1
Streptostachys asperifolia	GB01-2012	Paspaleae	Piot et al. 2017	-	MF563369.1
Themeda quadrivalvis	MSV350	Andropogoneae	Dunning et al. 2017	SRX2735032	KY707773.1
Themeda sp.	Saarela 1833	Andropogoneae	Burke et al. 2016	-	KU291484.1
Themeda triandra	AL01	Andropogoneae	Dunning et al. 2017	SRX2735038	KY707767.1
Thyridolepis xerophila	Saarela 1643 (CAN)	Paniceae	Burke et al. 2016	-	NC_030616.1
Tristachya humbertii	MSV1369	Tristachyideae	Piot et al. 2017	-	MF563368.1
Urochloa reptans	Morden 1221 (HAW)	Paniceae	Burke et al. 2016	-	NC_030617.1
Whiteochloa capillipes	Duvall s.n. (DEK)	Paniceae	Burke et al. 2016	-	NC_030618.1
Zea mays	SAMEA4040121	Andropogoneae	NCBI SRA Archive	ERR1462673	NC_001666.2

2.3.3. Genome-wide nuclear analysis

A novel approach was developed to automatically assemble full or partial coding sequences of nuclear genes from low-coverage genome scans. A genome-wide reference dataset of groups of co-orthologous genes at the Panicoideae subfamily level (i.e. all genes descended from a single gene in the most recent common ancestor of Panicoideae via a combination of speciation and duplication events) was retrieved from Dunning et al. (2017b). Genes potentially transferred from organelles to the nuclear genome were identified via BLAST (e-value threshold of 10^{-6}) using *Sorghum bicolor* organellar genomes as reference, and subsequently removed from this dataset. The final genome-wide reference dataset consisted of 11,313 groups of Panicoideae co-orthologs. The sequences of *S. bicolor* were extracted from this dataset and used as references for downstream analyses. These genes are descended from a single gene in the common ancestor of Panicoideae, but might be duplicated in some subgroups of Panicoideae or other grasses. Collapsing such duplicates allows extracting phylogenetically useful markers.

Gene models corresponding to each of the 11,313 references were assembled independently for each of the 98 grass species included here. First, raw genomic datasets were filtered using NGSQC Toolkit v.2.3.3 (Patel and Jain 2012) to retain only high-quality reads (i.e. > 80% of the bases with Phred quality score > 20), and to remove adaptor contamination and reads with ambiguous bases. The retained reads were subsequently trimmed from the 3' end to remove bases with Phred score < 20. The filtered genomic datasets were then mapped as single-end reads to the genome-wide reference using Bowtie2 v.2.3.2 (Langmead 2012) with default parameters. Consensus sequences were called based on variant call format (VCF) files from the read alignments using the *mpileup* function of Samtools v.1.5 (Li et al. 2009) implemented in a bash-scripted pipeline, modified from Chapter 4. IUPAC ambiguity codes were used for all variable sites. Sequences of all species and for all 11,313 genes were then concatenated to generate an initial supermatrix of 18,787,352 bp. Sites available for less than 70% of species were removed using trimAl v.1.4 (Capella-Gutiérrez et al. 2009), decreasing the proportion of missing data in the supermatrix from 74% to 26%, and a ML tree was inferred on the resulting 124,487 bp alignment with RAxML as described above.

In addition to the genome-wide dataset, eight individual nuclear markers previously used to infer grass phylogenies (GPWG 2001; Bomblies and Doebley 2005;

Doust et al. 2007; Christin et al. 2012b; Estep et al. 2012, 2014) were investigated, namely *aberrant panicle organization 1* (*apo1*), arogenate dehydrogenase (*arodeh*), the DELLA protein-encoding gene *dwarf 8* (*dwarf8*), *floricaula/leafy-like* (*floricaula*), *knotted 1* (*kn1*), phytochrome B (*phyB*), *retarded palea 1* (*rep1*) and granule-bound starch synthase 1 (GBSSI or *waxy*). Sequences for the putative C_3 sister group of Andropogoneae and the Arundinellinae *Garnotia stricta* var. *longiseta* were manually assembled by mapping reads to reference sequences in Geneious and calling a majority-rule consensus, as described in Besnard et al. (2018). Preliminary gene assembly for *Jansenella neglecta* revealed two divergent copies, of which one was very similar to the sequence of *J. griffithiana,* suggesting a hybrid origin (e.g. allopolyploid) of *J. neglecta*. However, this species was not included in the phylogenetic analyses because the relatively low sequencing depth prevented phasing the reads into alleles for the selected genes. The assembled sequences were aligned with additional data retrieved from NCBI nucleotide databases using MAFFT. Trees were inferred for each of the eight markers using MrBayes v.3.2.2 (Ronquist et al. 2012) with the GTR+G substitution model. Two parallel analyses consisting of four chains each were run for 40,000,000 generations. After verifying the convergence of the runs, the burn-in period was set to 50%, and a consensus tree was inferred for each gene using all the posterior trees for the parallel analyses.

2.3.4. Molecular dating

Divergence times were estimated for the nuclear and plastid datasets using a relaxed molecular clock as implemented in BEAST v.1.8.4 (Drummond and Rambaut 2007). To further reduce missing data in the nuclear dataset, sites available for less than 75% of species and species with more than 25% missing data were removed from the initial supermatrix. The final nuclear dataset included 66 species, with an alignment length of 32,501 bp and a proportion of missing data of 9%. The plastome alignment was reduced to coding sequences (57,239 bp) of the same 66 species for dating. The trees were time-calibrated by fixing the age of the split between the two main groups of grasses, the PACMAD and BOP clades, to 51.2 Ma (based on a previous analysis of nuclear datasets with different fossil calibration points; Prasad et al. 2011; Christin et al. 2014), using a normal distribution with standard deviation of 0.0001. This age represents the scenario based on macrofossils only, but we also report ages with the equivalent based on

disputed microfossils (82.4 Ma for the same node; Christin et al. 2014). The GTR+G substitution model was used, with the Yule model as speciation prior and a lognormal uncorrelated relaxed clock (Drummond et al. 2006). For each dataset, two or three MCMC chains were run in parallel for at least 160,000,000 generations. The runs were monitored using Tracer v.1.6 (Rambaut et al. 2013), checking for convergence and effective sample sizes >100 for all parameters. The burn-in period was set to the point of convergence of the runs (20%) and all trees sampled after that were combined. For each dataset, median ages were summarized on the maximum clade credibility tree.

2.3.5. Carbon isotopes and leaf anatomy

Photosynthetic types were retrieved from the literature (Osborne et al. 2014). The photosynthetic type of *J. griffithiana* was verified through analysis of carbon isotopes. Leaf fragments from the sequenced herbarium specimen were analysed using an ANCA GSL preparation module coupled to a Sercon 20-20 stable isotope ratio mass spectrometer (PDZ Europa, Cheshire, UK). Carbon isotopic ratios ($\delta^{13}C$, in ‰) were reported relative to the standard Pee Dee Belemnite (PDB). Values of $\delta^{13}C$ ranging from -33 to -24‰ are typical of C_3 plants, and values higher than -17‰ indicate that the plants grew using a C_4 pathway (O'Leary 1988).

Leaf anatomical phenotypes were characterized for members of Andropogoneae and their C_3 relatives, using data from the literature (Renvoize 1982a, 1982b, 1982c, 1985; Ueno 1995; Zuloaga et al. 2000; Christin et al. 2013; Watson et al. 1992). In addition, new leaf cross sections were prepared for the samples of *J. griffithiana* and *G. stricta* used for genome sequencing. A leaf fragment (ca. 2 cm) was rehydrated by warming the sample in dH_2O up to 60°C followed by immersion in 1% KOH overnight. The rehydrated fragment was then dehydrated through an ethanol series from 10% to 100% EtOH, with steps of 30 min. The leaf fragment was then resin-infiltrated with Technovit 7100 (Heraeus Kulzer GmbH, Wehrheim, Germany). Cross sections of 9 µm were obtained using a microtome (Leica RM 2245, Leica Biosystems Nussloch GmbH, Nussloch, Germany) and stained with Toluidine Blue O (Sigma-Aldrich, St. Louis, MO). Micrographs were obtained using an Olympus BX51 microscope coupled to an Olympus DP71 camera (Olympus Corporation, Tokyo, Japan). A number of qualitative and quantitative leaf characters related to the C_4 function were measured on the cross sections following Christin et al. (2013): number of bundle sheath layers, distance

between the centres of consecutive veins (interveinal distance), minimal distance between the bundle sheaths of consecutive veins (bundle sheath distance), fraction of the mesophyll plus bundle sheath area represented by the inner bundle sheath (% inner sheath area), presence/absence of distinctive cells (*sensu* Tateoka 1958; Renvoize 1982b), and localization of starch production.

2.3.6. Positive selection tests

To test for episodes of adaptive evolution of C_4 enzymes during different periods of the history of Andropogoneae, positive selection tests were conducted on alignments of five genes encoding proteins known to play important roles in the C_4 pathway (Hatch 1987), namely NADP-malate dehydrogenase (NADP-MDH), NADP-malic enzyme (NADP-ME), phosphoenolpyruvate carboxykinase (PCK), phosphoenolpyruvate carboxylase (PEPC) and pyruvate, phosphate dikinase (PPDK). Complete or partial coding sequences for the C_3 sister group of Andropogoneae and *G. stricta* were manually assembled, as described above. Additional sequences retrieved from NCBI and Washburn et al. (2017) were included, and the datasets were aligned using MAFFT. The 3rd positions of codons were used for phylogenetic inference to decrease biases due to adaptive evolution (Christin et al. 2012a). Phylogenetic trees were obtained using Bayesian inference with MrBayes as described above.

The trees were used to conduct positive selection analyses, after pruning C_4 species outside Andropogoneae to avoid selection signals in other C_4 groups, and enforcing the monophyly of Arundinellinae in one of the gene trees (PPDK). Removing genes from other C_4 taxa was necessary to avoid inflating the dN/dS estimated for the background branches or underestimating it in foreground branches by misidentifying C_4-specific genes in these other taxa that are not the focus of the present study. A number of codon models were optimized using *codeml* as part of PAML v. 4.9 (Yang 2008).The null model, assuming no selection, was compared to several branch-site models hypothesizing shifts in the selective pressure in some sets of foreground branches defined a priori: (1) the branch leading to Andropogoneae and its C_3 sister group (positive selection before the transition to C_4); (2) the branch leading to Andropogoneae (positive selection during the transition to C_4); and (3) the branches leading to each of the two main Andropogoneae groups Arundinellinae and Andropogoneae *s.s.* (positive selection just after the transition to C_4). Each model was

repeated with sustained selection from the selected branches to all descendants, and with a shift to relaxed selection instead of positive selection in foreground branches. The best model was selected using the Akaike Information Criterion (AIC), after verifying that it was significantly better than the null model (at a significance level of 5%) as assessed via a likelihood ratio test with a p-value adjusted for multiple testing using the Bonferroni correction.

The number of amino acid substitutions through time was assessed by estimating the branch lengths on the amino acid alignment while constraining the topology to that obtained on 3rd positions of codons. This was performed for the five core C_4 genes, using IQ-tree v.1.6.1 (Nguyen et al. 2015) with an automated selection of the model of protein sequence evolution.

2.4. Results

2.4.1. Nuclear and plastid trees

We inferred nuclear and plastid ML trees on a dataset consisting of 98 grass species. Both nuclear and plastid trees recovered most of the relationships among the major groups of grasses expected based on previous studies, with some notable exceptions (Figs 2.S1 and 2.S2). Among subfamilies, both the nuclear and plastome trees placed Aristidoideae as sister to the other PACMAD subfamilies, which is consistent with previous analyses of large nuclear datasets (Moreno-Villena et al. 2018), individual chloroplast markers (e.g. GPWG II 2012), some plastome partitions (Saarela et al. 2018), but not with the largest plastome analyses to date (Cotton et al. 2015; Burke et al. 2016; Saarela et al. 2018). Incongruence between the phylogenies recovered from plastid and nuclear genomes is observed within the well-sampled Panicoideae subfamily (Fig. 2.1), which points to different histories of the genomes, as recently suggested using a smaller dataset (Washburn et al. 2015, 2017). The C_3 genera *Jansenella* and *Chandrasekharania* form a strongly supported group sister to Andropogoneae, with a bootstrap value (BS) of 100 in both nuclear and plastome trees (Figs 2.S1 and 2.S2). This relationship was also highly supported in all trees for individual nuclear markers (Fig. 2.S3), although the group was paraphyletic in one case (*apo1*). Our data and analysis therefore provide strong evidence that the clade formed by the genera *Jansenella* and *Chandrasekharania* (hereafter Jansenelleae) is the C_3 lineage most closely related to the Andropogoneae grasses.

Within Andropogoneae, the group formed by the genera *Garnotia* and *Arundinella* (subtribe Arundinellinae) is sister to the Andropogoneae *s.s.* in all analyses, but occasionally paraphyletic with respect to Andropogoneae (Fig. 2.1; Fig. 2.S3). Short internal branches and low bootstrap support values within Andropogoneae *s.s.* are associated with high incongruence between nuclear and plastid trees, suggesting a complex history for the group, which might be related to a rapid radiation and frequent hybridization (Estep et al. 2014). In particular, nuclear and plastid trees identify different taxa as sister to the rest of Andropogoneae *s.s.* (*Arthraxon* in the nuclear tree, *Lasiurus* in the plastome tree; Figs 2.S1 and 2.S2).

2.4.2. Divergence time estimates and biogeography

The confirmation of the sister relationship between Jansenelleae and Andropogoneae allows refined divergence time estimates, including for the origin of C_4 photosynthesis in the group. Based on a secondary calibration considering only macrofossils, the nuclear and plastid datasets provided similar dates of divergence between Andropogoneae and its C_3 sister lineage at 21.3 (95% HPD = 16.5 – 26) Ma and 21.1 (95% HPD = 14.6 – 27.6) Ma, respectively (Fig. 2.1; Table 2.2). These dates would be pushed back to 33.9 and 33.5 Ma, respectively, if a microfossil dating scenario were followed. The first split within Andropogoneae was estimated at 19.2 (95% HPD = 14.9 – 23.8) Ma and 17.9 (95% HPD = 12.2 – 23.7) Ma for nuclear and plastid datasets, respectively (30.5 and 28.5 under a microfossil dating scenario).

We present a summary of the global distribution of species diversity and endemism of the Jansenelleae-Andropogoneae lineage in Figure 2.2. Of the three species in Jansenelleae, two are restricted to the Western Ghats of India and only *J. griffithiana* extends to other regions of mainland South East Asia, including Sri Lanka, Myanmar and Thailand. Andropogoneae s.s. and subtribe Arundinellinae have a wider distribution, but species diversity and endemism, relative to the TDWG boundaries, are also concentrated in India and South East Asia.

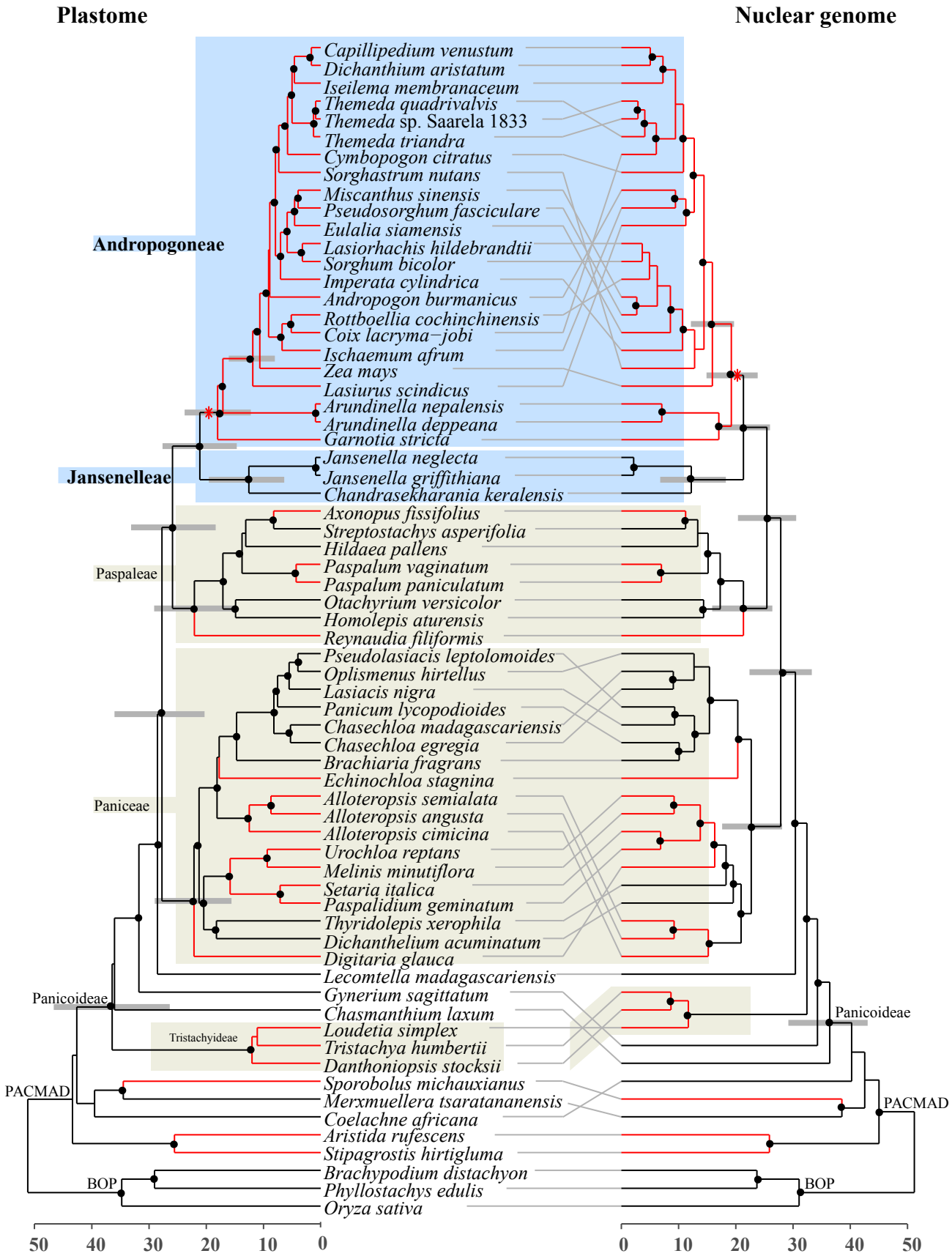

Plastome
Nuclear genome
Capillipedium venustum
Dichanthium aristatum
Iseilema membranaceum
Themeda quadrivalvis
Themeda sp. Saarela 1833
Themeda triandra
Cymbopogon citratus
Sorghastrum nutans
Miscanthus sinensis
Pseudosorghum fasciculare
Eulalia siamensis
Lasiorhachis hildebrandtii
Sorghum bicolor
Imperata cylindrica
Andropogon burmanicus
Rottboellia cochinchinensis
Coix lacryma−jobi
Ischaemum afrum
Zea mays
Lasiurus scindicus
Arundinella nepalensis
Arundinella deppeana
Garnotia stricta
Andropogoneae
Jansenella neglecta
Jansenella griffithiana
Chandrasekharania keralensis
Jansenelleae
Axonopus fissifolius
Streptostachys asperifolia
Hildaea pallens
Paspalum vaginatum
Paspalum paniculatum
Otachyrium versicolor
Homolepis aturensis
Reynaudia filiformis
Paspaleae
Pseudolasiacis leptolomoides
Oplismenus hirtellus
Lasiacis nigra
Panicum lycopodioides
Chasechloa madagascariensis
Chasechloa egregia
Brachiaria fragrans
Echinochloa stagnina
Alloteropsis semialata
Alloteropsis angusta
Alloteropsis cimicina
Urochloa reptans
Melinis minutiflora
Setaria italica
Paspalidium geminatum
Thyridolepis xerophila
Dichanthelium acuminatum
Digitaria glauca
Paniceae
Lecomtella madagascariensis
Gynerium sagittatum
Chasmanthium laxum
Loudetia simplex
Tristachya humbertii
Danthoniopsis stocksii
Tristachyideae
Panicoideae
Sporobolus michauxianus
Merxmuellera tsaratananensis
Coelachne africana
Aristida rufescens
Stipagrostis hirtigluma
PACMAD
Brachypodium distachyon
Phyllostachys edulis
Oryza sativa
BOP
50
40
30
20
10
0
0
10
20
30
40
50

Fig. 2.1. Time-calibrated trees of grasses based on (A) plastid and (B) genome-wide nuclear data. Divergence times were inferred from reduced 66-taxon datasets with columns with non-coding regions removed from plastomes (A) and columns with high amounts of missing data removed for nuclear genome sequences (B). The full 98-taxon maximum likelihood trees can be found in Figs 2.S1 and 2.S2 (Supplemental Information). The BOP–PACMAD split was calibrated at 51.2 Ma following Christin et al. (2014). Branches are coloured in red for C_4 and black for C_3 species, and the red star indicates the C_4 origin in Andropogoneae. Major taxonomic groups in Panicoideae are shaded. Dots on nodes indicate Bayesian posterior support ≥0.95.

Table 2.2. Divergence time estimates for selected lineages of grasses based on plastome and nuclear genome sequences[1].

Clade / Dataset	Plastome	Genome-wide nuclear markers
BOP crown	34.7 (24.8 – 45.6)	31 (17.8 – 44.3)
PACMAD crown	43.4 (34.6 – 51.1)	45 (37.8 – 51.2)
Panicoideae crown	36.4 (26.3 – 46.6)	36.4 (29.2 – 43)
Jansenelleae / Andropogoneae split	21.1 (14.6 – 27.6)	21.3 (16.5 – 26)
Andropogoneae crown	17.9 (12.2 – 23.7)	19.2 (14.9 – 23.8)
Andropogoneae *s.s.* crown	11.9 (8 – 16)	15.9 (12.1 – 19.7)

[1] Values are median ages in million years ago (Ma), with 95% HPD intervals in parentheses.

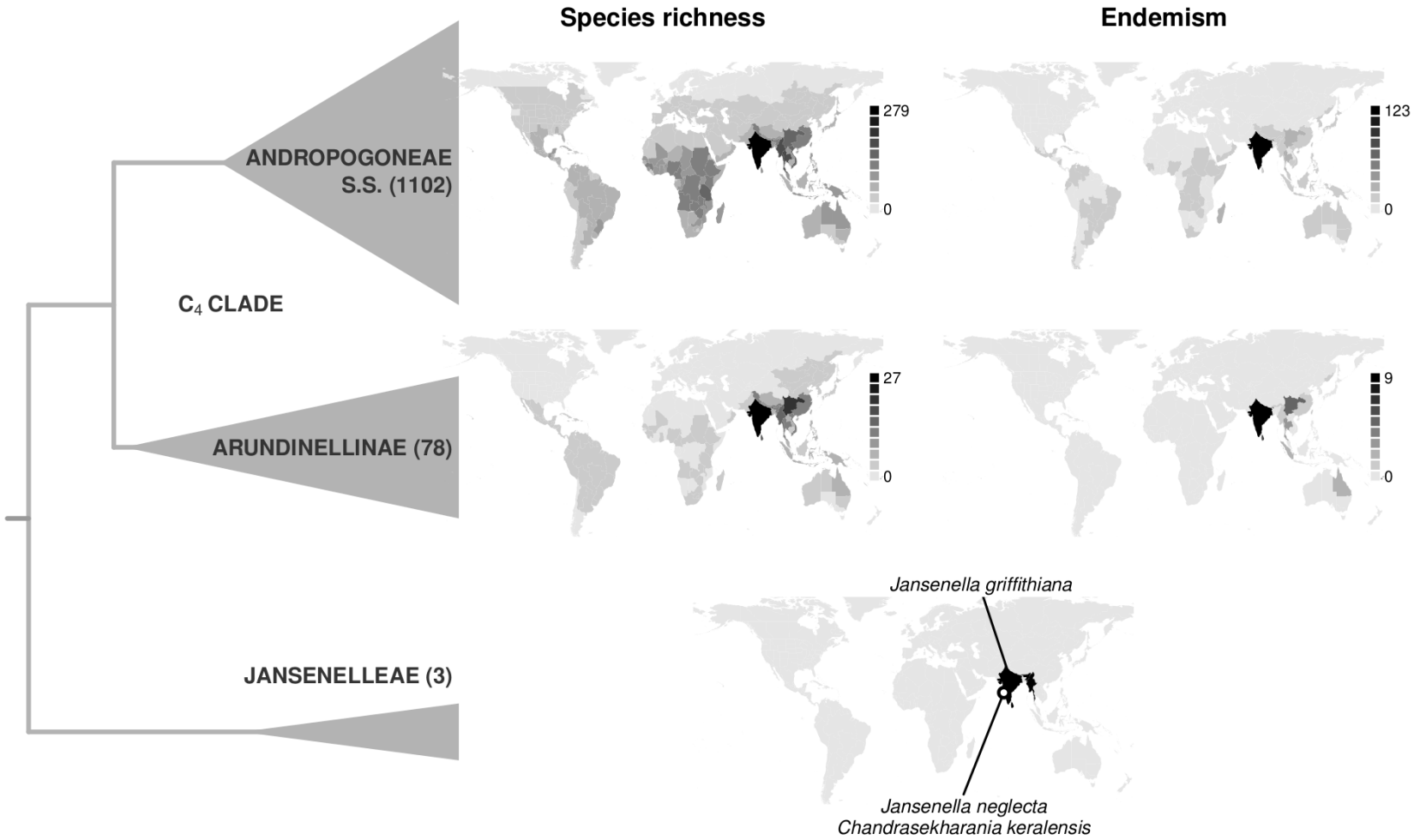


Fig. 2.2. Distribution and endemism of Andropogoneae and Jansenelleae lineages. Species numbers given are relative to World Geographical Scheme for Recording Plant Distributions (TDWG) level-3 botanical regions.

2.4.3. Anatomical changes during the early diversification of Andropogoneae

Anatomical and biochemical characters linked to C_4 photosynthesis were assessed based on the literature and new measurements for various Andropogoneae and their C_3 relatives (Table 2.3). Our carbon isotope analysis confirmed that *J. griffithiana* is a C_3 plant ($\delta^{13}C$ = -27.28 ‰). Its leaf anatomy is typical of C_3 grasses, with two layers of bundle sheath cells (Fig. 2.3A), which contradicts previous reports (Metcalfe 1960; Türpe 1970). As with other C_3 species, it has a large distance between consecutive bundle sheaths, and no sign of minor veins or distinctive cells (Fig. 2.3B). In addition, the proportion of the leaf occupied by the inner bundle sheath falls within the range observed for other C_3 grasses (Christin et al. 2013). The leaf anatomy of *G. stricta* was similar to that previously reported for Arundinellinae (Renvoize 1982c, 1986; Watson 2017). Its veins are surrounded by a single bundle sheath and are separated by a large number of mesophyll cells (Fig. 2.3B). Multiple distinctive cells separate the veins, and staining suggests starch production in both bundle sheaths and distinctive cells (Fig. 2.3B). Similar anatomical structures are observed in other Arundinellinae (Renvoize 1982c, 1986; Watson 2017), but also in the genus *Arthraxon* (Ueno 1995), which is an Andropogoneae *s.s.* representative that diverged early from the rest of the group (e.g. GPWG II 2012; Estep et al. 2014). By contrast, most Andropogoneae *s.s.* lack distinctive cells and decrease the distance between consecutive veins via the proliferation of minor veins (Fig. 2.3B; Table 2.3). The distribution of these traits on the phylogeny suggests either a switch between minor veins and distinctive cells at the base of Andropogoneae *s.s.*, or independent origins of these characters.

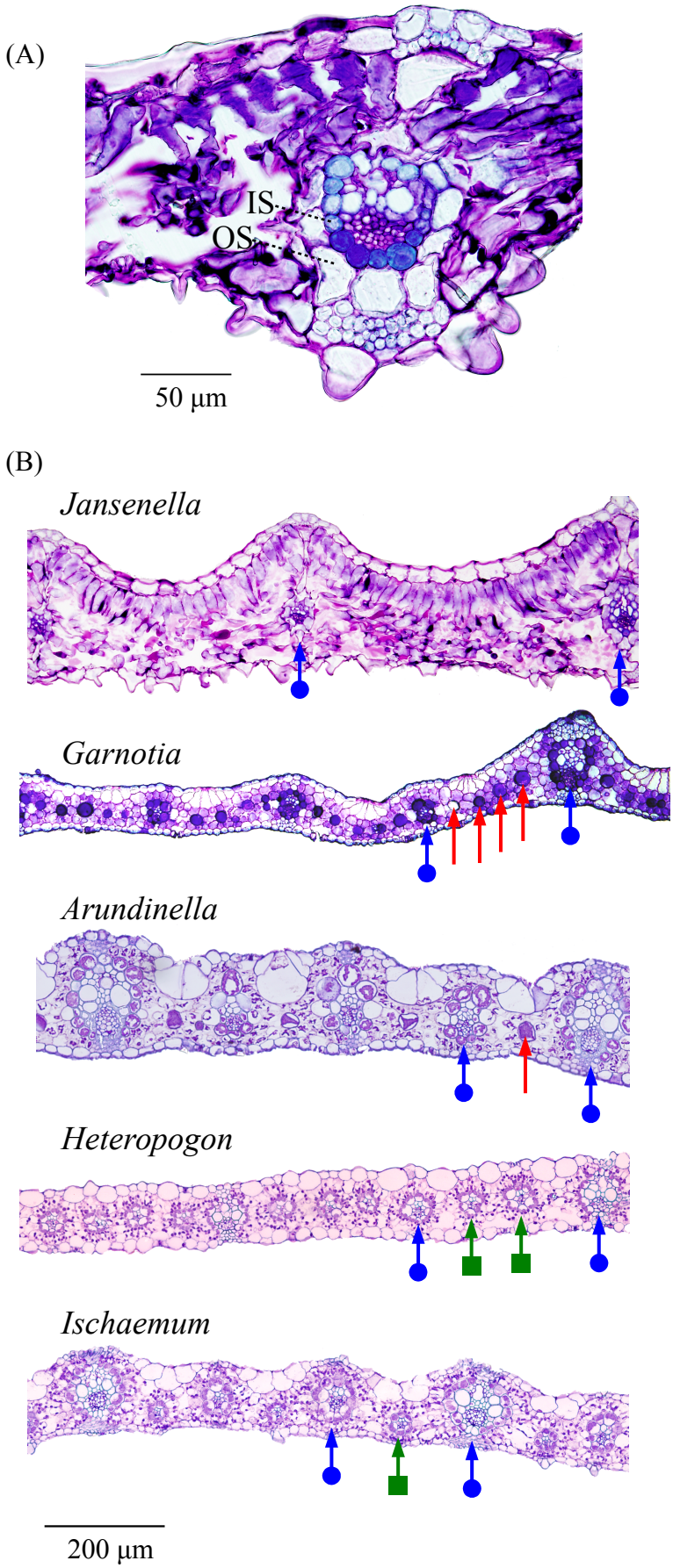


Fig. 2.3. Leaf transverse sections of representatives of Jansenelleae and Andropogoneae. (A) *Jansenella griffithiana*. OS = outer bundle sheath; IS = inner bundle sheath; (B) *Jansenella griffithiana*, *Garnotia stricta*, *Arundinella nepalensis*, *Heteropogon contortus* and *Ischaemum afrum*. The latter three pictures are from Christin et al. (2013). Blue, red and green arrows indicate major veins, distinctive cells and minor veins, respectively.

Table 2.3. Leaf anatomical measurements of traits associated with C_4 photosynthesis in *Jansenella griffithiana* and representatives of Andropogoneae.

Lineage/Species	Bundle sheath layers	Interveinal distance (μm)	% Inner Sheath Area	Bundle sheath distance (μm)	Distinctive cells	Starch in BSC
Jansenelleae						
Jansenella griffithiana (C_3)	2	471	0.01	403	A	A
Arundinellinae						
Arundinella nepalensis (C_4)[1]	1	215	0.27	101	P	P
Garnotia stricta (C_4)	1	191	0.11	187	P	P
Andropogoneae s.s.						
Arthraxon sp. (C_4)[2,3]	1	-	-		P	P
Chrysopogon pallidus (C_4)[1]	1	112	0.23	29	A	P
Heteropogon contortus (C_4)[1]	1	80	0.21	32	A	P
Ischaemum afrum (C_4)[1]	1	109	0.24	52	A	P
Sorghum halepense (C_4)[1]	1	119	0.20	53	A	P

A = absent, P = present; [1] Data extracted from Christin et al. (2013); [2] Data extracted from Watson et al. (1992 onwards); [3] Data extracted from Ueno (1995).

2.4.4. Positive selection in C_4 enzymes

Phylogenetic trees for genes encoding C_4 enzymes inferred from 3rd positions of codons were compatible with genome-wide trees (Fig. 2.S4). In all cases, Jansenelleae were sister to Andropogoneae, in which Arundinellinae and Andropogoneae *s.s.* represent the first split, with one exception (PPDK), in which Arundinellinae is paraphyletic (Fig. 2.S4). Lineage-specific duplications are observed within Andropogoneae *s.s.* and Arundinellinae species for NADP-ME, and only in Andropogoneae *s.s.* for NADP-MDH (Fig. 2.S4), as previously reported (Rondeau et al. 2005; Christin et al. 2009a; Wang et al. 2009).

The inferred trees were used to track episodes of adaptive evolution in Andropogoneae, independently for each gene. In all five core C_4 genes analysed, the best model inferred a shift of selective pressures in Andropogoneae which was sustained until the present (Table 2.4). For two genes (NADP-MDH and PPDK), the shift occurred at the base of Andropogoneae, and the model assuming a shift to positive selection was significantly better than the model assuming a shift to relaxed selection. In the three other genes (NADP-ME, PCK and PEPC), the shift of selective pressures

occurred independently in Arundinellinae and Andropogoneae *s.s.*; in these cases, a shift of selective pressures was statistically supported, but models assuming a shift to positive or relaxed selection performed equally well (Table 2.4).

In the model assuming shifts to positive selection, the Bayes Empirical Bayes approach (BEB) identified 25 amino acid sites of PEPC under different selective pressures in Andropogoneae. At all of these sites, there is variation in the amino acids observed in Arundinellinae and Andropogoneae *s.s.* species (Fig. 2.S4), and in most cases they differ from the C_3 outgroups. None of these 25 sites correspond to the 12 previously identified as convergently selected in multiple grass lineages (Christin et al. 2007). However, these 12 sites contained substitutions in all of the Andropogoneae species, except for *G. stricta*, in which four of these sites were not modified when compared to the C_3 sister group. A total of 29 sites of NADP-ME were identified by the BEB approach to be under different selective pressures in Andropogoneae. Interestingly, the amino acids at these sites varied among genes of Andropogoneae (Fig. 2.S4), suggesting that mutations accumulated gradually during the diversification of the group. To visualize the amount of amino acid substitutions during different periods of Andropogoneae history, we estimated branch lengths from amino acid sequences after excluding C_4 species outside of Andropogoneae. Overall, numerous substitutions occurred on PEPC, PPDK and to some extent on NADP-ME at the base of Andropogoneae, and increased rates compared to non-C_4 genes were sustained throughout Andropogoneae (Figs 2.4 and 2.5). By contrast, bursts of amino acid substitutions on NADP-MDH occurred mainly in the Andropogoneae *s.s.*, while increased number of substitutions on PCK were restricted to some species within both Arundinellinae and Andropogoneae *s.s.* (Figs 2.4 and 2.5). The same patterns were observed when C_4 species outside of Andropogoneae were included in the analyses (Fig. 2.S5). Increased rates of amino acid substitution on all five genes characterize most C_4 branches, as was observed in the Andropogoneae, which highlights the highly convergent nature of C_4 evolution in grasses.

Table 2.4. Positive selection tests in three scenarios of C_4 enzyme adaptation in the group including Jansenelleae and Andropogoneae[1].

			Branch site models											
Scenarios of adaptive evolution			**Single episode** (Preceding C_4 evolution)				**Single episode** (During C_4 evolution, at the base of Andropogoneae)				**Two episodes** (During C_4 evolution, at the base of Arundinellinae and Andropogoneae *s.s.*)			
Gene[2] (Enzyme)	n species	Site model M1a (null model)	Internal branch only		Sustained selection		Internal branch only		Sustained selection		Internal branch only		Sustained selection	
			BSA	BSA1	BSA	BSA1	BSA	BSA1	BSA	BSA1	BSA	BSA1	BSA	BSA1
nadpmdh-1P1 (NADP-MDH)	32	45.50	47.50	49.50	13.12	5.87	44.78	46.78	9.74	**0.00***	47.5	49.50	13.93	2.38
nadpme-1P4 (NADP-ME)	51	248.58	250.58	252.58	57.41	59.41	250.58	252.58	2.33	4.19	250.58	252.58	**0.00***	1.84
pck-1P1 (PCK)	39	65.40	67.40	69.40	15.68	17.68	67.40	69.40	2.61	4.61	67.40	69.40	**0.00***	2.00
ppc-1P3 (PEPC)	56	516.68	518.68	520.68	44.49	46.49	479.67	481.67	17.01	19.01	483.59	483.97	**0.00***	2.00
ppdk-1P2 (PPDK)	28	59.88	61.88	63.88	20.11	19.70	61.88	63.88	3.80	**0.00***	58.06	59.50	6.36	3.66

[1] dAIC values relative to the best-fit model for each gene are shown. The best-fit model is highlighted in bold with an asterisk. Two hypotheses of enzyme adaptation were tested for each scenario, the first assuming positive selection in the internal branch leading to the ancestor of the group specified in the scenarios, the second assuming sustained selection, which includes the ancestor plus all descendant branches. For each hypothesis, two branch-site models were used to test for a relaxation of purifying selection (BSA), and potential positive selection (BSA1).

[2] Gene notation following Moreno-Villena et al. (2018)

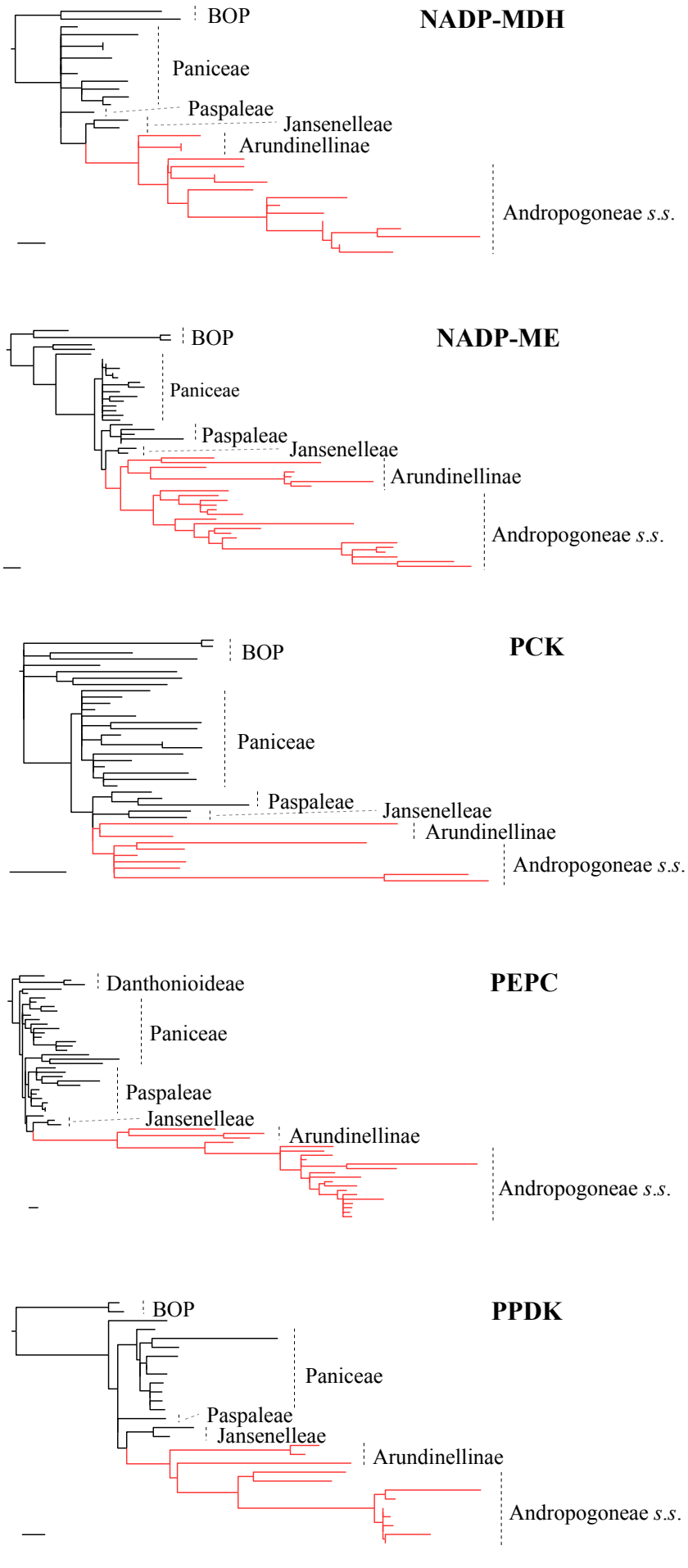


Fig. 2.4. Protein trees with topologies constrained to that obtained using 3rd positions of codons for genes encoding five core C_4 enzymes. Scale bar = 0.01 amino acid substitutions per site. Branches are coloured in red for C_4 and black for C_3 accessions.

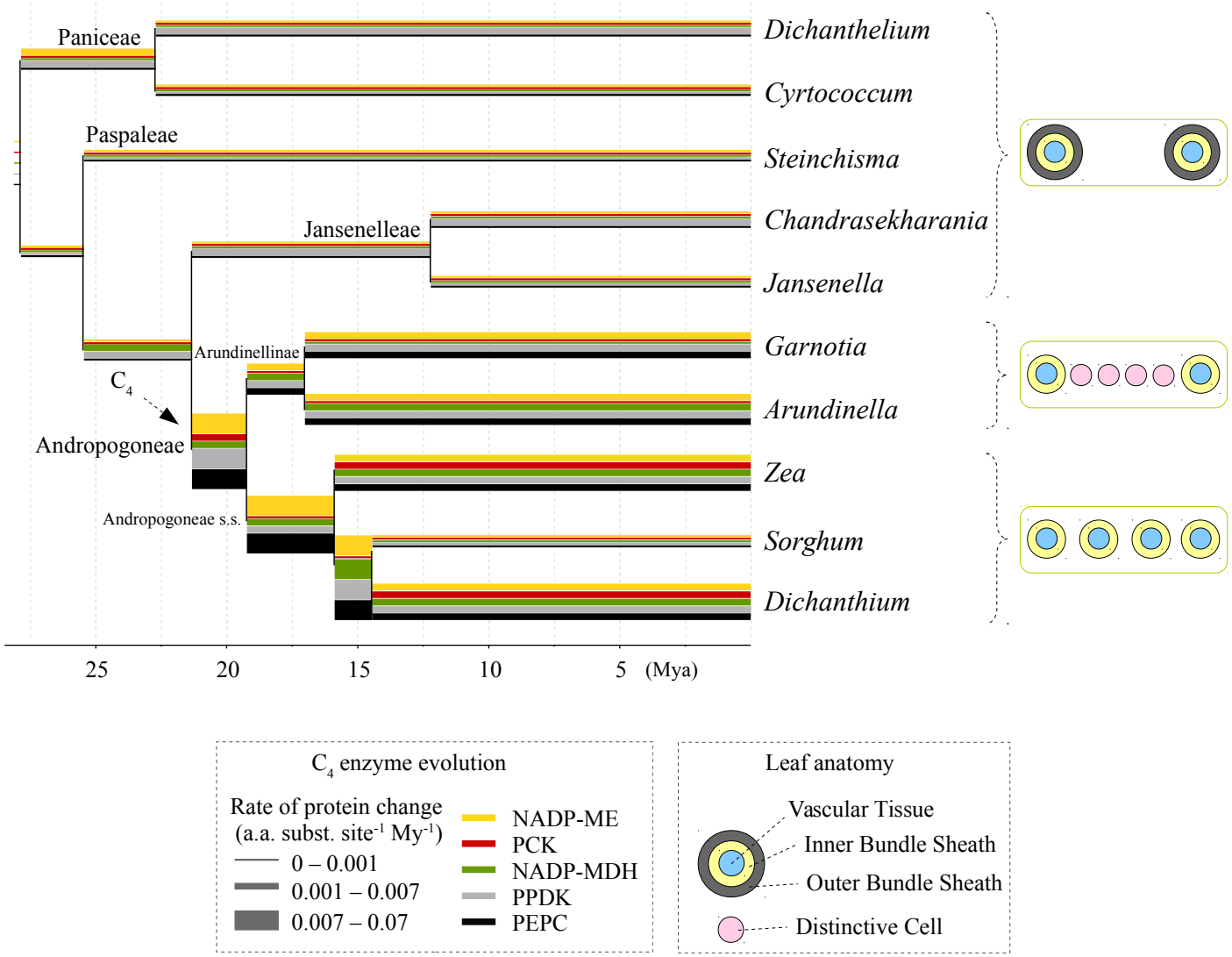


Fig. 2.5. Changes in protein sequence and leaf anatomy of C_4 components in the Andropogoneae grasses. A time-calibrated tree of Panicoideae is presented, with selected non-C_4 species outside Andropogoneae. Branch thickness is proportional to the rate of protein change and colours represent different C_4 enzymes. A simplified transverse section of the leaf is presented on the right, with colours representing the different tissues.

2.5. Discussion

2.5.1. A single origin of the new C_4 physiology followed by continued anatomical changes

In previous grass phylogenies, Andropogoneae formed a large clade entirely composed of C_4 species, and its closest known C_3 relatives belonged to a different group in which there were multiple independent C_4 lineages (GPWG II 2012). The branch leading to Andropogoneae was therefore long, preventing the precise inference of changes leading to C_4 evolution in this group. We confirm here that *Jansenella* and *Chandrasekharania* form the sister group of Andropogoneae, both based on plastome phylogenies and on markers spread across the nuclear genomes. This, combined with a distinctive morphology, supports their recognition as a separate tribe, Jansenelleae (Appendix 1). We further confirm that the group is C_3, as previously suggested (Türpe 1970; Renvoize 1985, 1986), providing a shorter branch connecting the last known C_3 ancestor of Andropogoneae (most recent ancestor shared with Jansenelleae) and the first split within the group. The anatomy of *Jansenella* is typical of C_3 grasses, with a large distance between consecutive veins, a double bundle sheath and no minor veins or distinctive cells (Fig. 2.3). In addition, the genes encoding C_4-related enzymes from *Jansenella* and *Chandrasekharania* are similar to those of other C_3 grasses, with no traces of positive selection or increased rates of amino acid replacements (Figs 2.4, 2.S4 and 2.S5; Table 2.4). We therefore conclude that the last common ancestor of Jansenelleae and Andropogoneae was a typical C_3 plant, with the anatomical and genetic characteristics common to all PACMAD grasses (Christin et al. 2013; Emms et al. 2016; Moreno-Villena et al. 2018). The changes responsible for the emergence of a C_4 pathway therefore happened after the divergence between Andropogoneae and Jansenelleae. Previous studies comparing C_3 and C_4 anatomical traits or genomes typically sampled only a few Andropogoneae species, preventing assigning changes to different phases of C_4 evolution (Christin et al. 2013; Emms et al. 2016; Huang et al. 2017), as enabled here thanks to our denser species sampling.

The comparison of anatomical types suggests multiple modifications during the early diversification of Andropogoneae. All species from this group have a single bundle sheath (Renvoize 1982), which is ontogenetically equivalent to the inner sheath of C_3 grasses (i.e. mestome sheath; Dengler et al. 1985). The large distance between

consecutive veins, as observed in *Jansenella* (Table 2.3), is reduced in Arundinellinae by the insertion of one or multiple distinctive cells, where Rubisco can be segregated (Fig. 2.3; Dengler and Dengler 1990; Sinha and Kellogg 1996). While these distinctive cells are shared by some Andropogoneae *s.s.* (Ueno 1995), most use a different strategy to reduce the distance between consecutive veins, which consists of the proliferation of minor veins (Table 2.3; Lundgren et al. 2014). Distinctive cells and minor veins have similar developmental patterns (Dengler et al. 1996), and the former could be precursors of the latter, in which case minor veins could represent the specialization of ancestral distinctive cells after the split of Andropogoneae *s.s.* from Arundinellinae. Alternatively, the ancestral state of the group could be minor veins that later degenerated in Arundinellinae and some Andropogoneae, or else these specializations evolved multiple times during the early diversification of the group. In all cases, the phylogenetic distribution of distinctive cells and minor veins shows that changes following the initial transition to C_4 led to diverse anatomical solutions for the effective segregation of biochemical reactions.

2.5.2. Modifications of C_4 enzymes occurred throughout the diversification of Andropogoneae

It is accepted that the emergence of a C_4 pathway requires the co-option of multiple enzymes already existing in the C_3 ancestor via their massive upregulation (Hibberd and Covshoff 2010; Moreno-Villena et al. 2018). This is followed by adaptation of their kinetics for the new catalytic context through numerous amino acid replacements (Blasing et al. 2002; Tausta et al. 2002; Christin et al. 2007; Huang et al. 2017). Positive selection tests conducted here for multiple C_4-encoding genes from Andropogoneae and other grasses confirm that the evolution of C_4 genes in this group involved numerous adaptive modifications of the coding sequences (Table 2.4; Fig. 2.S4; Christin et al. 2007, 2009; Wang et al. 2009; Huang et al. 2017). The key enzyme of the C_4 pathway, PEPC, underwent convergent changes in numerous groups of grasses, and most were shared between *Arundinella* and Andropogoneae *s.s.* (Christin et al. 2007). However, only a fraction of these changes are also observed in *Garnotia stricta*, indicating that the enzyme underwent adaptive changes both before and after the split among the major lineages of Andropogoneae (Fig. 2.S4), and codon models did not favour positive selection at the base of the whole clade (Table 2.4). The assumption of adaptive

evolution on the branch leading to Andropogoneae is supported for genes encoding NADP-MDH, NADP-ME, and PPDK (Table 2.4), and a comparison of branch lengths shows increased rates of sustained amino acid replacements in genes for these three C_4 enzymes at the base of Andropogoneae (Fig. 2.4). Our analyses therefore confirm that massive changes happened at the base of Andropogoneae, but models assuming that positive selection persisted after early episodes of adaptive evolution are strongly favoured for at least two genes (Table 2.4). In addition, increased rates of sustained amino acid replacements are observed on many branches within the group. We conclude that important alterations of enzymes for the initial build-up of a C_4 cycle at the base of Andropogoneae were followed by continued adaptation throughout the diversification of the group.

While some enzymes participate in all biochemical variants of the C_4 cycle (Kanai and Edwards 1999), the identity of the enzyme(s) responsible for the decarboxylation of CO_2 in the bundle sheath varies among C_4 lineages (Prendergast et al. 1987; Sage et al. 2011). Our analyses concordantly indicate that the decarboxylating enzyme PCK underwent rounds of amino acid replacements only in some derived groups within Andropogoneae (Figs 2.4 and 2.S4), without evidence of positive selection at the base of the whole group (Table 2.4). This conclusion was reached previously (Christin et al. 2009b) and supports later additions of a PCK-catalyzed decarboxylation reaction in some of the Andropogoneae (Gutierrez et al. 1974; Walker et al. 1997; Wingler et al. 1999). However, our data also indicate that NADP-ME, which is the main decarboxylating enzyme in all Andropogoneae, similarly acquired its C_4 properties relatively late in the history of the group. Again, the best model assumed adaptive evolution throughout Andropogoneae (Table 2.4). Genes for NADP-ME were duplicated independently in Andropogoneae *s.s.*, *Garnotia* and *Arundinella*, and amino acid replacements are especially prevalent in one of the copies in each group (Figs 2.4 and 2.S4; Christin et al. 2009a). These observations point to independent adaptation of the enzyme kinetics, but the expression patterns also likely evolved independently in Andropogoneae *s.s.* and Arundinellinae. Indeed, modifications of the promoter regions allowing the C_4-specific binding of a transcription factor are restricted to one of the Andropogoneae *s.s.* duplicates that fulfils the C_4 function (Borba et al. 2018), which evolved after the split from Arundinellinae. We therefore hypothesize that the common ancestor of the Andropogoneae performed a C_4 cycle based on several decarboxylating enzymes relatively abundant in many C_3 grasses (Moreno-Villena et al. 2018), with

some amino acid changes in the other C_4 enzymes. Further modifications, which canalized the use of NADP-ME, added a PCK shuttle, or improved the action of PEPC, PPDK and NADP-MDH, happened later during the diversification of the group, so that its numerous C_4 species represent a diversity of realizations of the C_4 pathway. Similar conclusions were reached for small groups that evolved the C_4 trait recently (Dunning et al. 2017a), but we show here for the first time that the continuous adaptation of the C_4 trait can be sustained over long evolutionary periods, leaving traces even within one of the largest C_4 groups.

2.5.3. C_4 physiology evolved during the early Miocene in Andropogoneae

Besides inferring the changes underlying C_4 evolution in Andropogoneae, our genome-wide, time-calibrated phylogenies encompassing a diversity of Andropogoneae and their closest C_3 relatives shed new light on the age and geographic origin of C_4 photosynthesis in the group. Our molecular dating estimated the split between Jansenelleae and Andropogoneae at roughly 21 Ma, with the first split within Andropogoneae at 18-19 Ma. While older ages would be inferred if disputed microfossils dates are considered (see Results), these dates represent the interval in which C_4 most likely evolved in this group. Reconstructing the ancient biogeography of Andropogoneae is complicated by their diversity and presumably numerous dispersals across large distances, but India represents the centre of diversity of Andropogoneae *s.s.* (already noted by Hartley 1958) and Arundinellinae as well as Jansenelleae (Fig. 2.2; Bor 1955; Nair et al. 1982; Yadav et al. 2010). We conclude that C_4 photosynthesis likely originated in this lineage on the Indian subcontinent in the Early Miocene. Once the three species of Jansenelleae occur in open habitats (Bor 1955; Nair et al. 1982; Yadav et al. 2010), the transition likely happened in full-light conditions, which favour the C_4 type in warm regions (Osborne and Freckleton 2009). Also, the occurrence of at least *J. griffithiana* in regularly burning grasslands (Shilla and Tiwari 2015) suggests that fire may already have played a structuring role before the transition. Increasing seasonality (Sage et al. 2012) and warming may have provided the selective impetus for the transition to C_4 in the group; the Asian monsoon system was probably established by ~24 Ma but intensified over the Miocene (Guo et al. 2002; Clift 2006; Clift et al. 2008), while global temperatures were high through the Early and Middle Miocene,

terminating with global cooling and regional drying of Asia ~15 Ma (Zachos et al. 2001; Clift 2006; Srivastava et al. 2018).

The contrast between the sister groups Jansenelleae and Andropogoneae is striking. While the former has only three extant species, two of them restricted to small regions of India, the latter encompasses roughly 1,200 species spread around the world, many of which are dominant in savanna ecosystems (Hulbert 1988; Solbrig 1996; Bond et al. 2003; Kellogg 2015). This difference is partially explained by the divergence of photosynthetic types, but the expansion of C_4 grasslands happened 7-15 Ma after C_4 originated in Andropogoneae (Edwards et al. 2010), and increased diversification occurred only in some of its subclades (Spriggs et al. 2014). While the initial C_4 trait might have played the role of a key innovation broadening the niche of early Andropogoneae (Lundgren et al. 2015), the later diversification and dominance of some subgroups, their rapid dispersal across large distances (Dunning et al. 2017b) and into different ecosystems (Watcharamongkol et al. 2018) were likely enabled by the acquisition of additional attributes. Traits only partially related or entirely unrelated to C_4 photosynthesis, such as frequent allopolyploidy, herbivore resistance and fire tolerance have previously been used to explain the success of some Andropogoneae (Stebbins 1975; Bond et al. 2003; Edwards et al. 2010; Visser et al. 2012; Estep et al. 2014; Forrestel et al. 2014; Ripley et al. 2015; Linder et al. 2018). We suggest that the diversity of C_4 phenotypes revealed here might also contribute to variation among Andropogoneae. For instance, the addition of a PCK shuttle, which happened recurrently in some derived Andropogoneae (Figs 2.4 and 2.S4), is predicted to increase tolerance to fluctuating light conditions (Bellasio and Griffiths 2014; Wang et al. 2014). Other anatomical and biochemical variations observed here might alter the hydraulic efficiency and growth rates of the different Andropogoneae. Overall, we conclude that, because of continuous adaptive reinforcement following a key physiological transition, descendants of a lineage sharing the derived trait should not all be considered as functionally equivalent.

2.6. Conclusions

We confirmed a rare C_3 lineage from the Indian subcontinent, Jansenelleae, as sister to the Andropogoneae grasses. This opens new avenues for comparative analysis of C_4 evolution, which were explored here. The C_4 pathway in Andropogoneae most likely

evolved in the Early Miocene between roughly 21 and 18 Ma, and many adaptive changes in C_4 enzymes happened during this 3-My period, while many more occurred during the next 18 million years of sustained adaptation. The group including Andropogoneae apparently originated on the Indian subcontinent, and the evolutionary diversification of the C_4 phenotype after its origin might have been associated with the spread of Andropogoneae into novel niches and to different regions of the globe, contributing to the success of this emblematic group of savanna grasses.

2.7. Acknowledgements

We thank Jacob Washburn for providing assembled transcriptomes for this study, Heather Walker for mass spectrometry analysis, Pierre Solbès (EDB lab), Luke Dunning, Jill Olofsson and Daniel Wood for bioinformatic support, Hans-Joachim Esser (Munich Botanical Gardens) for providing a herbarium sample, and Simone de Padua Teixeira and Lamiaa Munshi for suggestions on the leaf anatomy preparation. M.E.B. is supported by the Brazilian Research Council (CNPq) through a 'Science without Borders' scholarship (grant number 201873/2014-1). J.H. and G.B. received support from the French excellence projects Labex CEBA (ANR-10-LABX-25-01) and Labex TULIP (ANR-10-LABX-0041). This work was performed within the framework of the PhyloAlps project, whose sequencing was funded by France Génomique (ANR-10-INBS-09-08). P.A.C. is funded by a Royal Society University Research Fellowship (URF120119). MRM was supported by NSF grants DEB-11456884 and DEB-1457748 to EAK.

2.8. Appendix 1

Jansenelleae Voronts. & E.A.Kellogg tr. nov.
Type: *Jansenella* Bor, Kew Bull. 1955: 96. 1955.
Included genera: *Chandrasekharania, Jansenella*

Diagnosis:
Jansenelleae differ from Andropogoneae by their awned lower lemma (lower lemma is muticous in the Andropogoneae, or very rarely awned or absent).

Description:
Delicate annuals, rooting at the lower nodes, with ascending culms. Ligule shortly membranous. Leaf blades membranous, lanceolate. Inflorescence shortly branched, appearing capitate, sometimes with secondary branching. Spikelets paired, each spikelet pair with one short and one long pedicel, the pedicel disarticulating at the apex so one spikelet is a dispersal unit (upper floret also a dispersal unit in *Jansenella*). Sessile and pedicelled spikelets similar, laterally compressed, 2-flowered. Glumes 2, apically acuminate to shortly awned. Lower glume separated from the rest of the spikelet by an elongated rachilla internode. Lower floret sterile, staminate, or bisexual. Lower lemma shortly awned, either entire (*Jansenella*) or awn arising between two erose apical lobes (*Chandrasekharania*). Lower palea present. Upper floret bisexual. Upper lemma awned from a bidentate apex, but variable in its shape and indumentum: either with two hair tufts, twisted dehiscent awn arising between long-acuminate lobes (*Jansenella*) or without hair tufts, short straight awn arising between two erose apical lobes (*Chandrasekharania*). Stamens 3. Grain ellipsoid; hilum punctiform.

Leaf anatomy:
Outer and inner bundle sheaths present. More than four cells between consecutive veins. No distinctive cells. Starch storage in the chlorenchyma.

Distribution:
India (Maharashtra, Chhattisgarh, Odisha, Karnataka, Kerala, Tamil Nadu, Meghalaya); Sri Lanka; Myanmar (Bago); Thailand (Peninsular).

Notes:

The awned lower lemma appears to be a synapomorphy for the Jansenelleae. Only one species in the Andropogoneae sensu stricto has been recorded as having an awned lower lemma: *Microstegium somae* (Hayata) Ohwi (fide Hsu 1975; Shouliang and Phillips 2006). *Garnotia* in the Arundinellinae (*sensu* Kellogg 2015) is unlike the rest of the tribe with only a single floret, with a lemma which is sometimes awned, and cannot be meaningfully compared to the lower lemma of the two-flowered Panicoideae.

2.9. Supporting Information

Fig. 2.S1. Maximum likelihood tree based on complete plastomes. Red and black branches are C_4 and C_3 species, respectively. Bootstrap support values are shown on nodes (values < 50% were omitted).

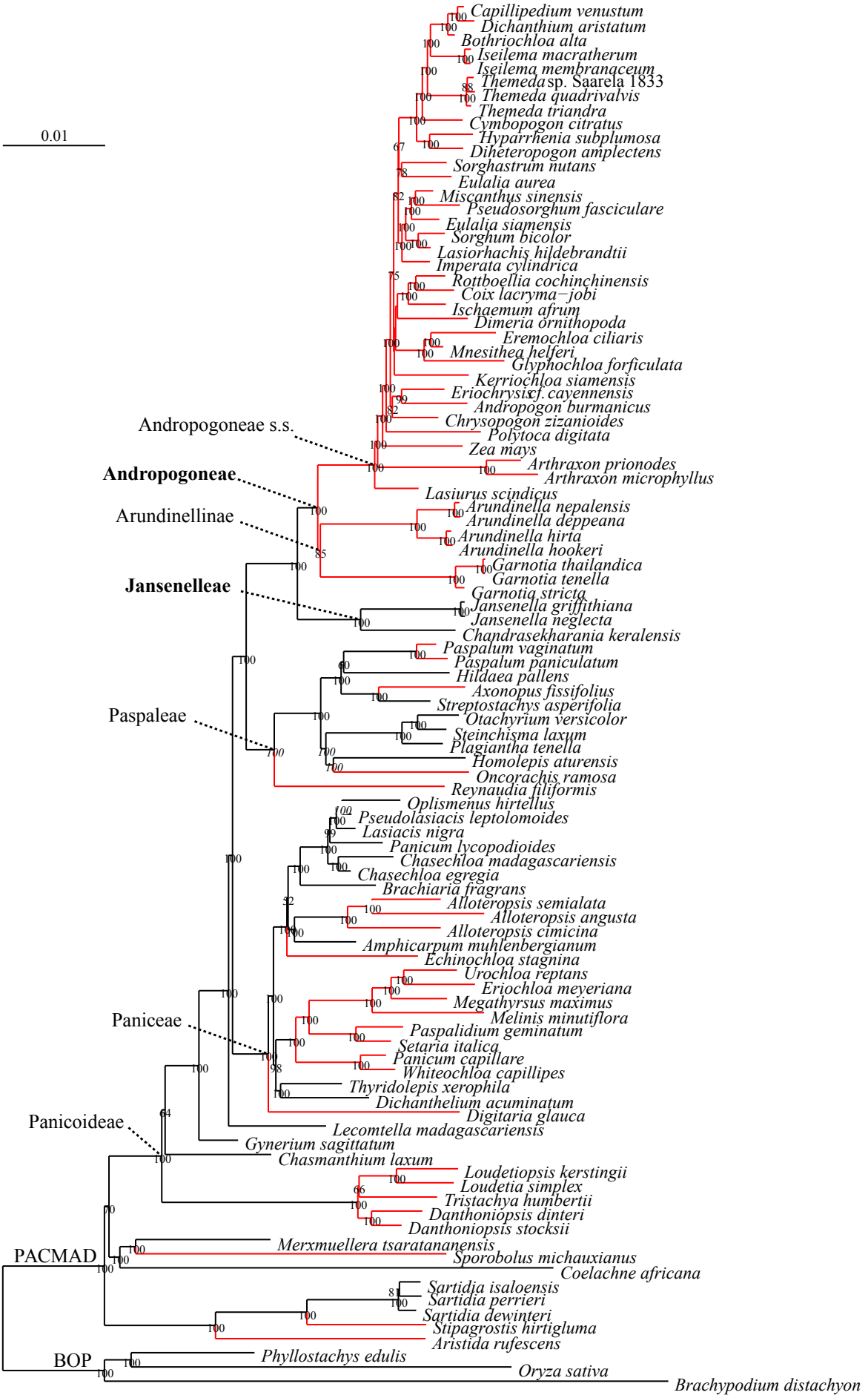

0.01
Capillipedium venustum
Dichanthium aristatum
Bothriochloa alta
Iseilema macratherum
Iseilema membranaceum
Themeda sp. Saarela 1833
Themeda quadrivalvis
Themeda triandra
Cymbopogon citratus
Hyparrhenia subplumosa
Diheteropogon amplectens
Sorghastrum nutans
Eulalia aurea
Miscanthus sinensis
Pseudosorghum fasciculare
Eulalia siamensis
Sorghum bicolor
Lasiorhachis hildebrandtii
Imperata cylindrica
Rottboellia cochinchinensis
Coix lacryma−jobi
Ischaemum afrum
Dimeria ornithopoda
Eremochloa ciliaris
Mnesithea helferi
Glyphochloa forficulata
Kerriochloa siamensis
Eriochrysis cf. cayennensis
Andropogon burmanicus
Chrysopogon zizanioides
Polytoca digitata
Zea mays
Arthraxon prionodes
Arthraxon microphyllus
Lasiurus scindicus
Arundinella nepalensis
Arundinella deppeana
Arundinella hirta
Arundinella hookeri
Garnotia thailandica
Garnotia tenella
Garnotia stricta
Jansenella griffithiana
Jansenella neglecta
Chandrasekharania keralensis
Paspalum vaginatum
Paspalum paniculatum
Hildaea pallens
Axonopus fissifolius
Streptostachys asperifolia
Otachyrium versicolor
Steinchisma laxum
Plagiantha tenella
Homolepis aturensis
Oncorachis ramosa
Reynaudia filiformis
Oplismenus hirtellus
Pseudolasiacis leptolomoides
Lasiacis nigra
Panicum lycopodioides
Chasechloa madagascariensis
Chasechloa egregia
Brachiaria fragrans
Alloteropsis semialata
Alloteropsis angusta
Alloteropsis cimicina
Amphicarpum muhlenbergianum
Echinochloa stagnina
Urochloa reptans
Eriochloa meyeriana
Megathyrsus maximus
Melinis minutiflora
Paspalidium geminatum
Setaria italica
Panicum capillare
Whiteochloa capillipes
Thyridolepis xerophila
Dichanthelium acuminatum
Digitaria glauca
Lecomtella madagascariensis
Gynerium sagittatum
Chasmanthium laxum
Loudetiopsis kerstingii
Loudetia simplex
Tristachya humbertii
Danthoniopsis dinteri
Danthoniopsis stocksii
Merxmuellera tsaratananensis
Sporobolus michauxianus
Coelachne africana
Sartidia isaloensis
Sartidia perrieri
Sartidia dewinteri
Stipagrostis hirtigluma
Aristida rufescens
Phyllostachys edulis
Oryza sativa
Brachypodium distachyon
Andropogoneae s.s.
Andropogoneae
Arundinellinae
Jansenelleae
Paspaleae
Paniceae
Panicoideae
PACMAD
BOP

Fig. 2.S2. Maximum likelihood tree based on genome-wide nuclear sequences. Red and black branches are C_4 and C_3 species, respectively. Bootstrap support values are shown on nodes (values < 50% were omitted).

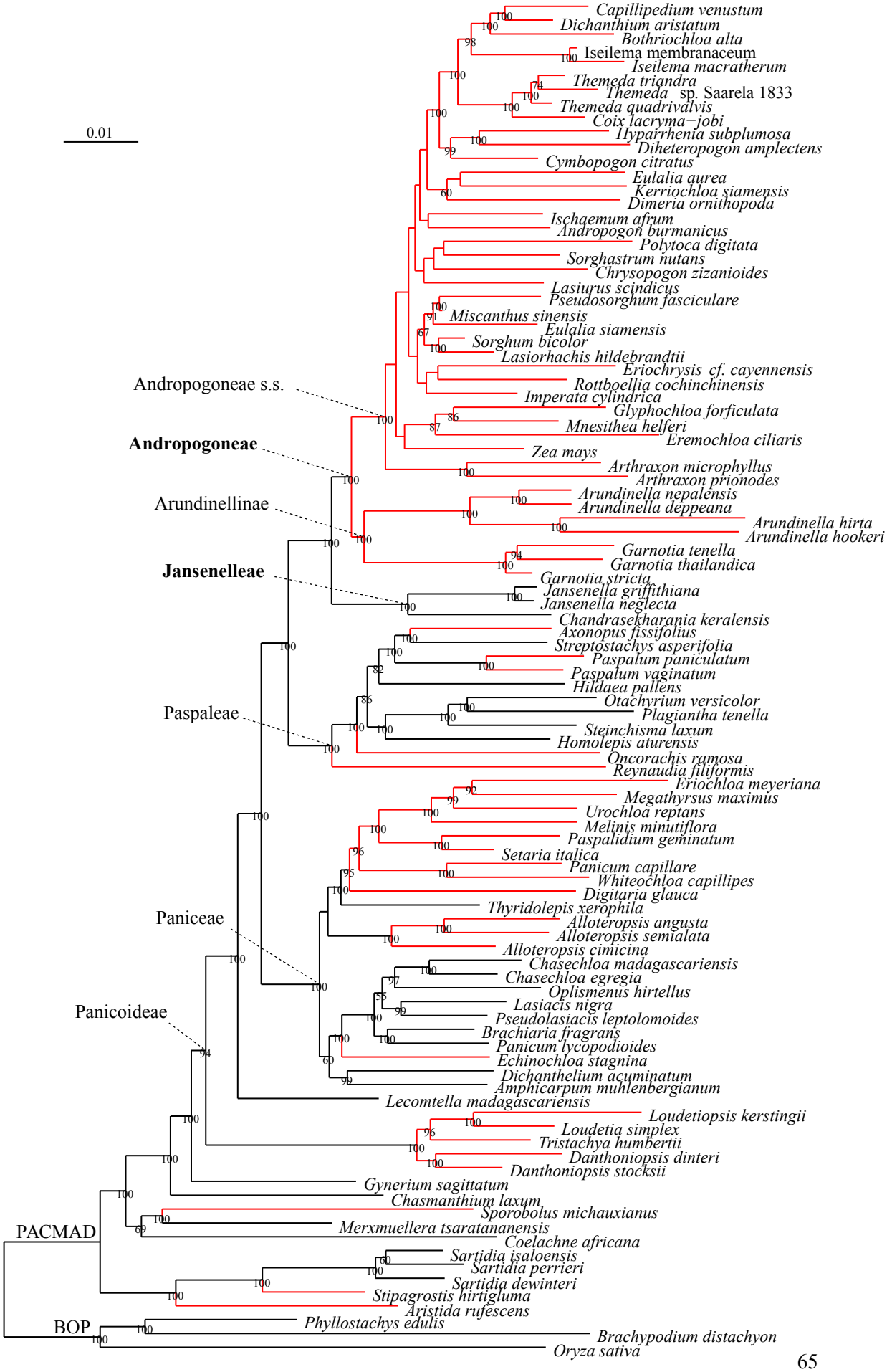
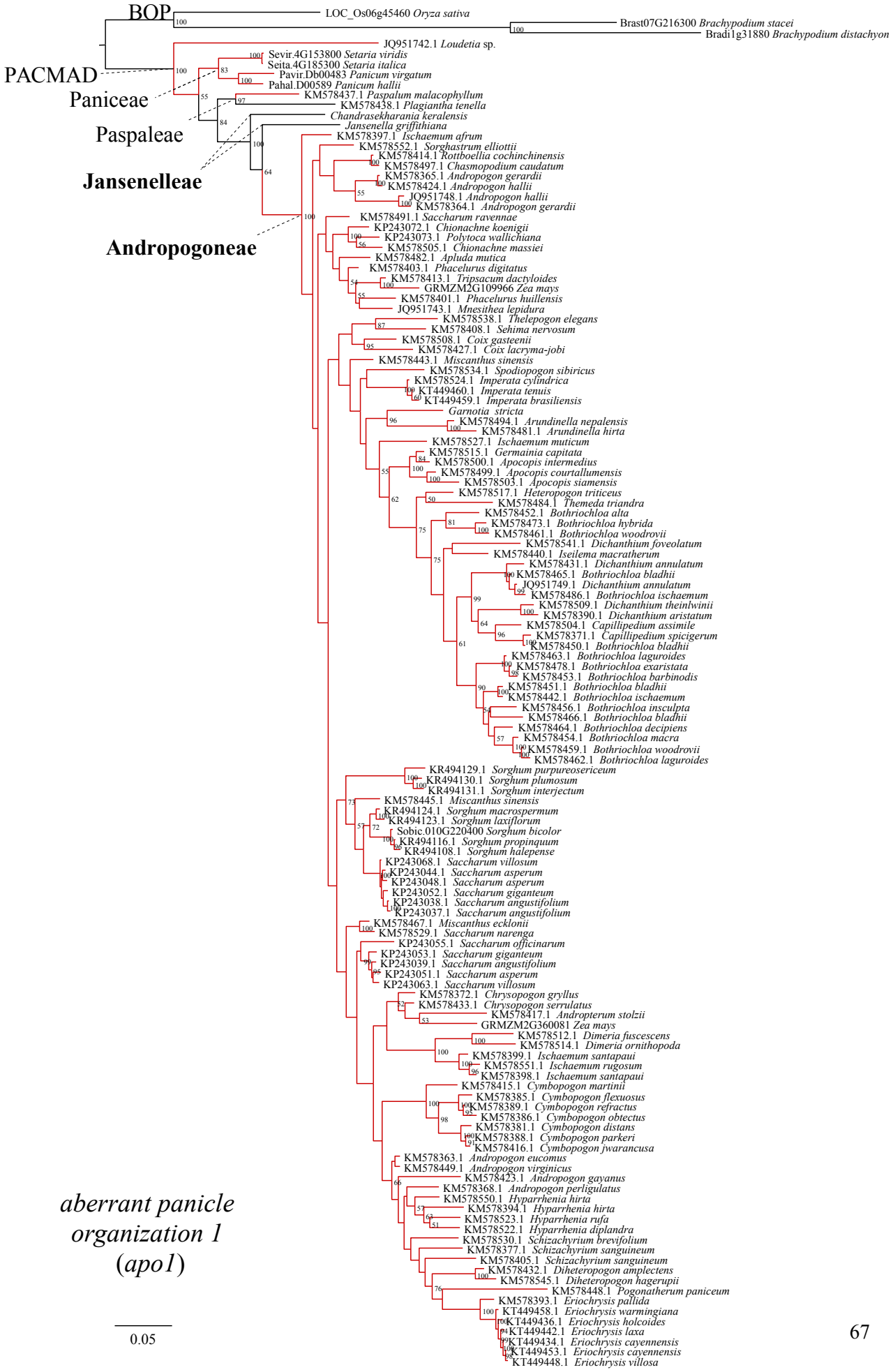
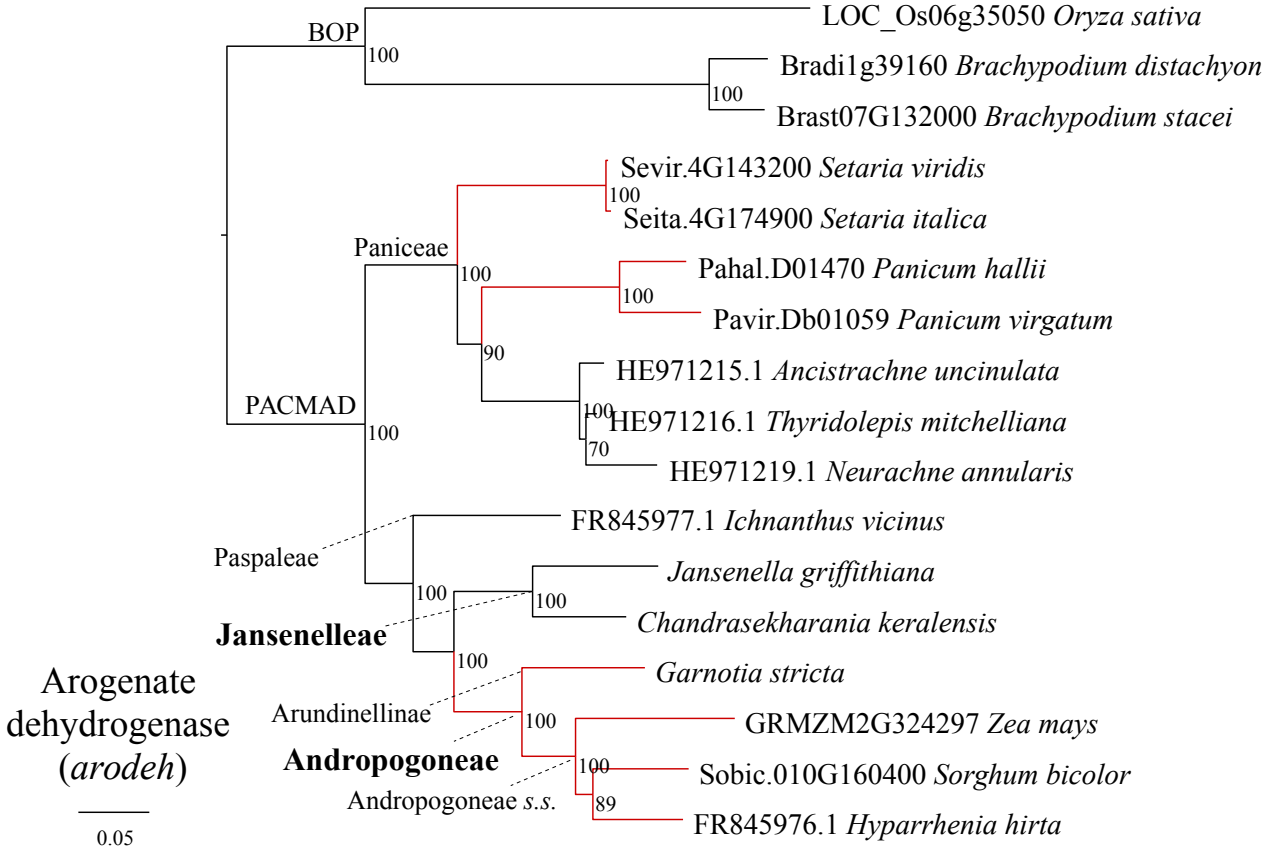
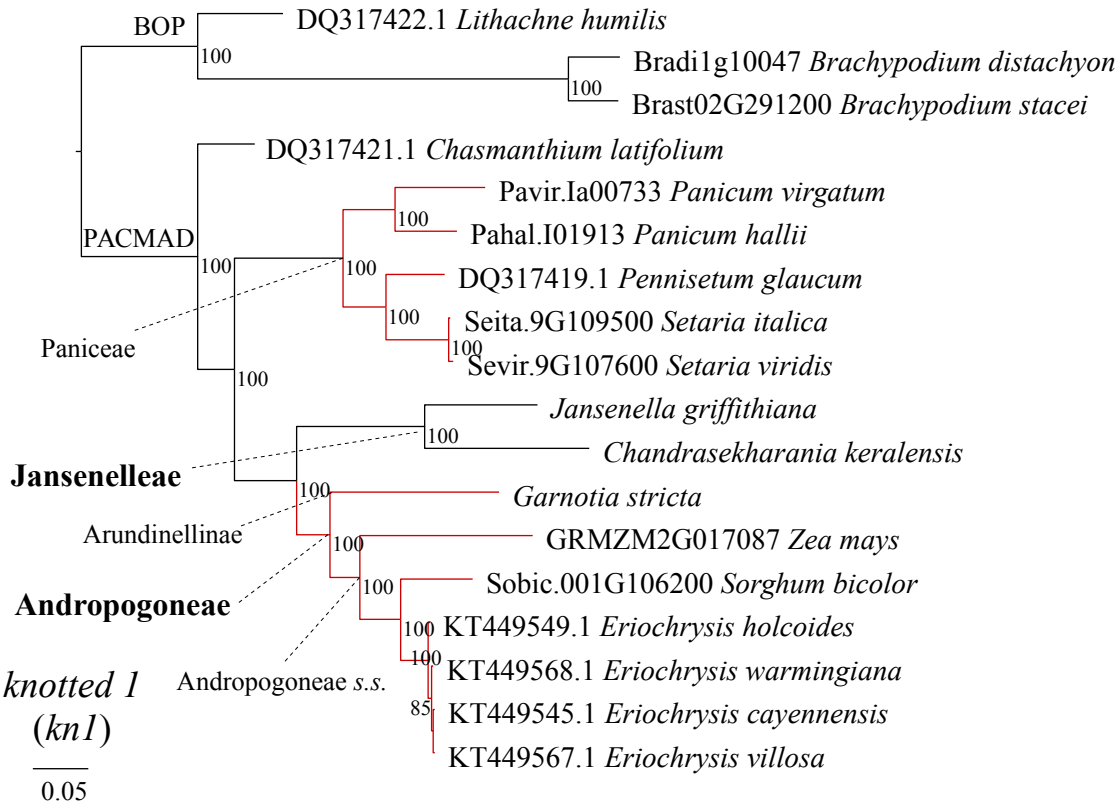

0.01
Capillipedium venustum
Dichanthium aristatum
Bothriochloa alta
Iseilema membranaceum
Iseilema macratherum
Themeda triandra
Themeda sp. Saarela 1833
Themeda quadrivalvis
Coix lacryma−jobi
Hyparrhenia subplumosa
Diheteropogon amplectens
Cymbopogon citratus
Eulalia aurea
Kerriochloa siamensis
Dimeria ornithopoda
Ischaemum afrum
Andropogon burmanicus
Polytoca digitata
Sorghastrum nutans
Chrysopogon zizanioides
Lasiurus scindicus
Pseudosorghum fasciculare
Miscanthus sinensis
Eulalia siamensis
Sorghum bicolor
Lasiorhachis hildebrandtii
Eriochrysis cf. cayennensis
Rottboellia cochinchinensis
Imperata cylindrica
Glyphochloa forficulata
Mnesithea helferi
Eremochloa ciliaris
Zea mays
Arthraxon microphyllus
Arthraxon prionodes
Arundinella nepalensis
Arundinella deppeana
Arundinella hirta
Arundinella hookeri
Garnotia tenella
Garnotia thailandica
Garnotia stricta
Jansenella griffithiana
Jansenella neglecta
Chandrasekharania keralensis
Axonopus fissifolius
Streptostachys asperifolia
Paspalum paniculatum
Paspalum vaginatum
Hildaea pallens
Otachyrium versicolor
Plagiantha tenella
Steinchisma laxum
Homolepis aturensis
Oncorachis ramosa
Reynaudia filiformis
Eriochloa meyeriana
Megathyrsus maximus
Urochloa reptans
Melinis minutiflora
Paspalidium geminatum
Setaria italica
Panicum capillare
Whiteochloa capillipes
Digitaria glauca
Thyridolepis xerophila
Alloteropsis angusta
Alloteropsis semialata
Alloteropsis cimicina
Chasechloa madagascariensis
Chasechloa egregia
Oplismenus hirtellus
Lasiacis nigra
Pseudolasiacis leptolomoides
Brachiaria fragrans
Panicum lycopodioides
Echinochloa stagnina
Dichanthelium acuminatum
Amphicarpum muhlenbergianum
Lecomtella madagascariensis
Loudetiopsis kerstingii
Loudetia simplex
Tristachya humbertii
Danthoniopsis dinteri
Danthoniopsis stocksii
Gynerium sagittatum
Chasmanthium laxum
Sporobolus michauxianus
Merxmuellera tsaratananensis
Coelachne africana
Sartidia isaloensis
Sartidia perrieri
Sartidia dewinteri
Stipagrostis hirtigluma
Aristida rufescens
Phyllostachys edulis
Brachypodium distachyon
Oryza sativa
Andropogoneae s.s.
Andropogoneae
Arundinellinae
Jansenelleae
Paspaleae
Paniceae
Panicoideae
PACMAD
BOP

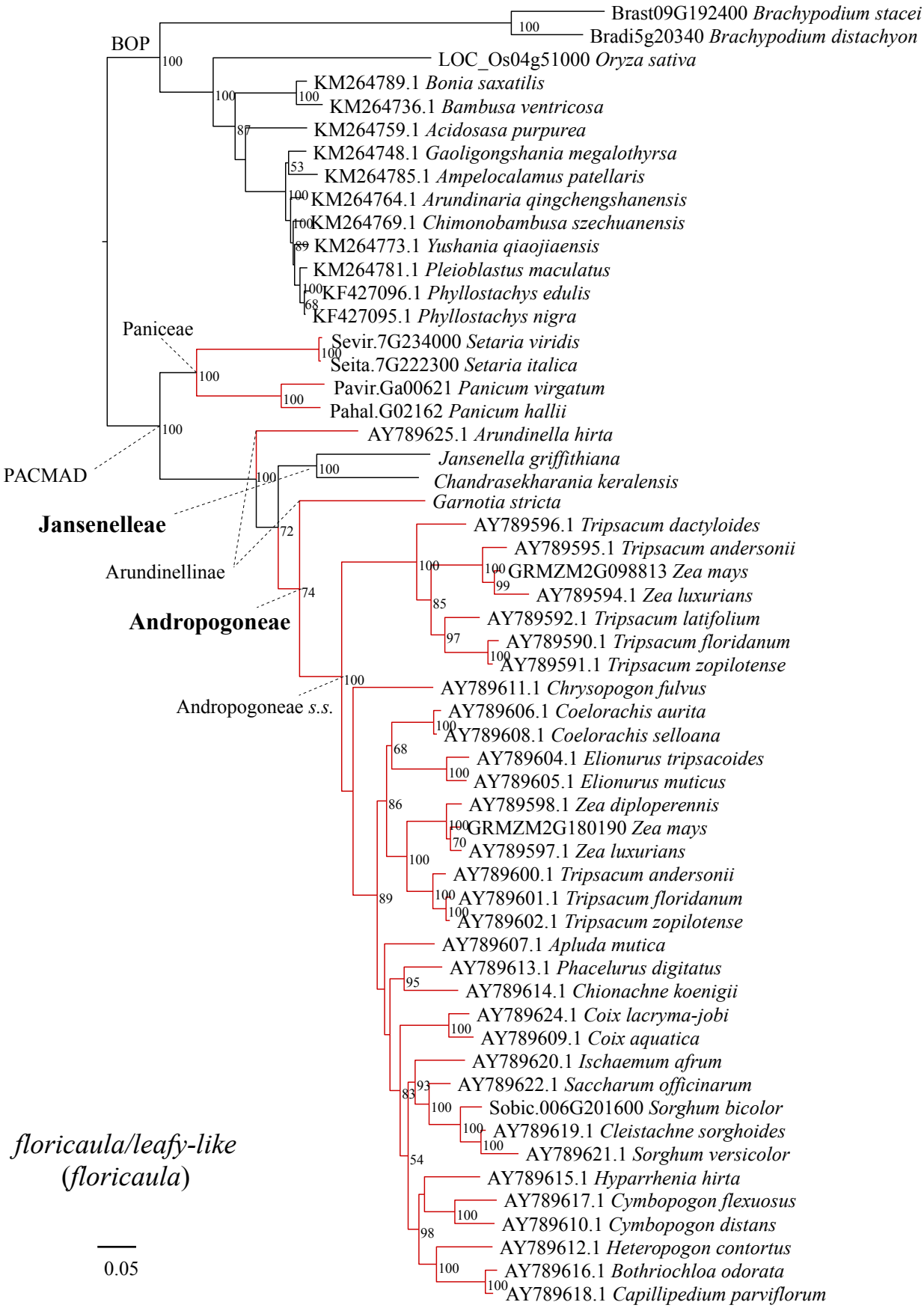
Fig. 2.S3. Bayesian trees based on individual nuclear markers. Red and black branches are C_4 and C_3 species, respectively. Posterior probabilities are shown near nodes (values < 50% were omitted).

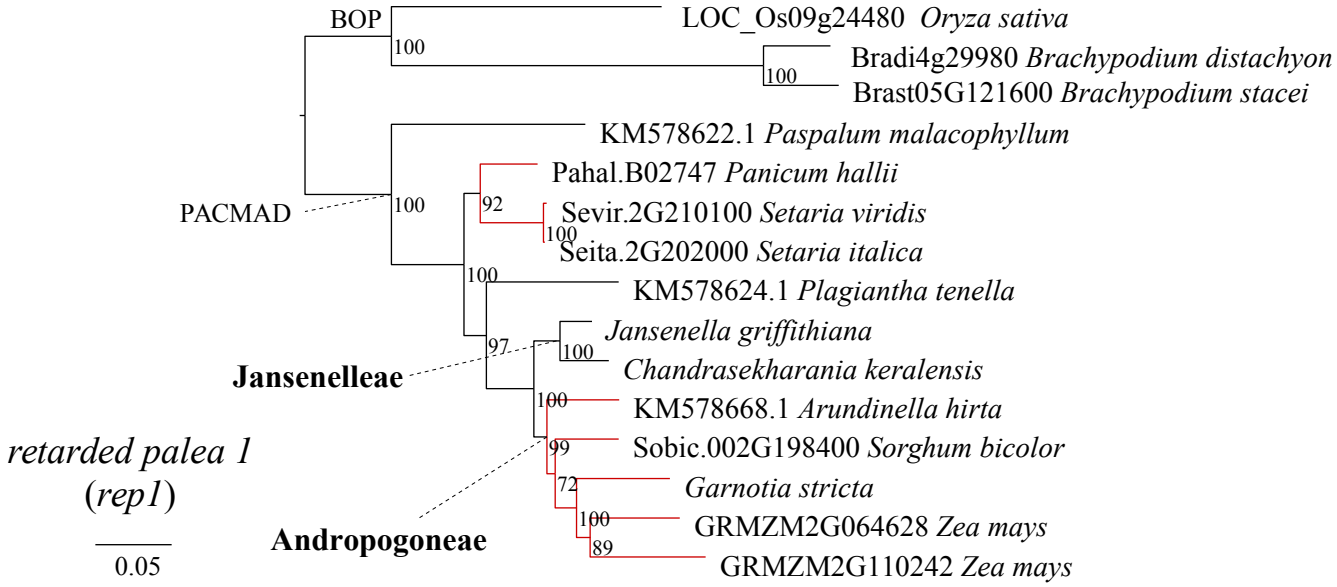


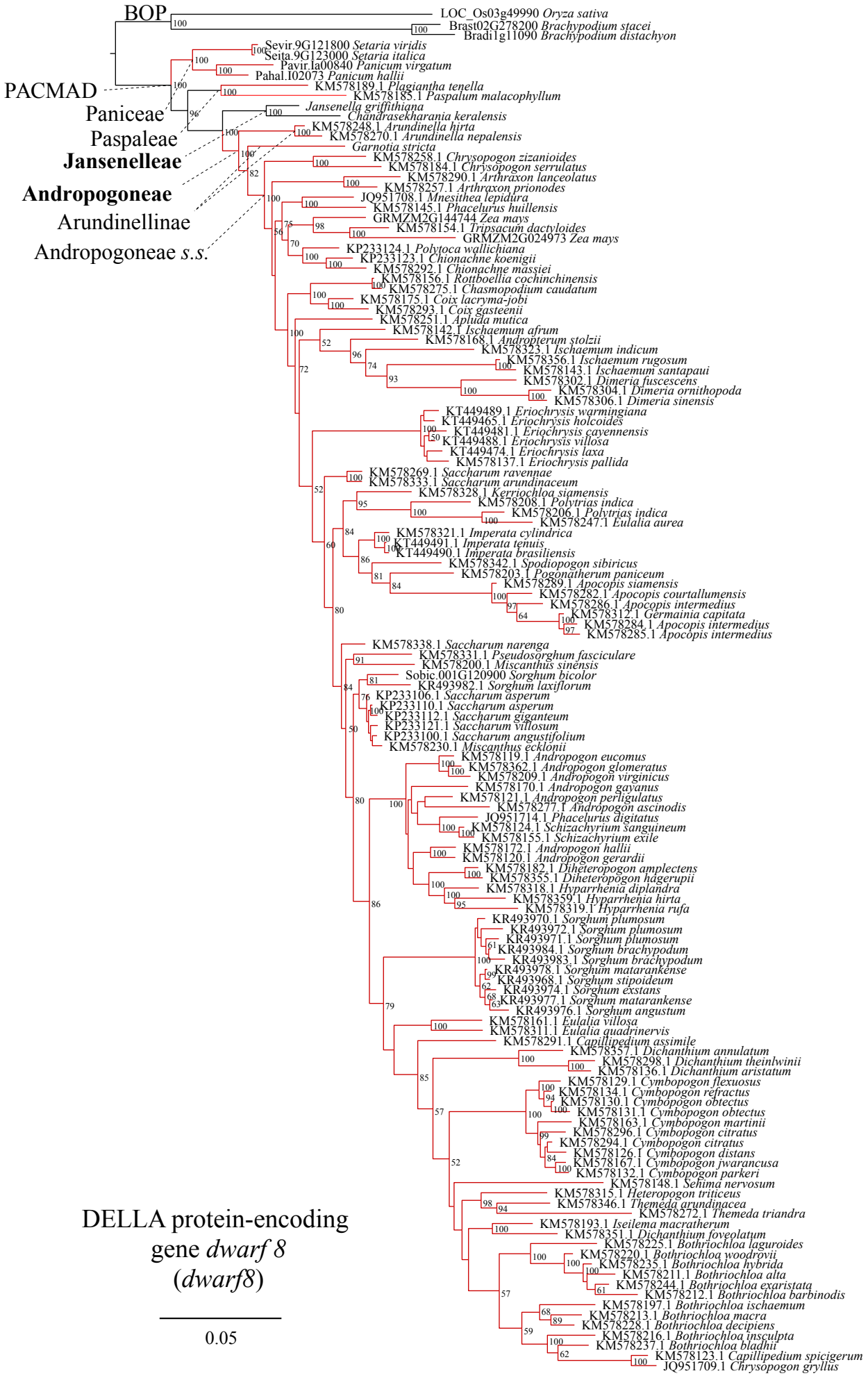
aberrant panicle organization 1 (apo1)

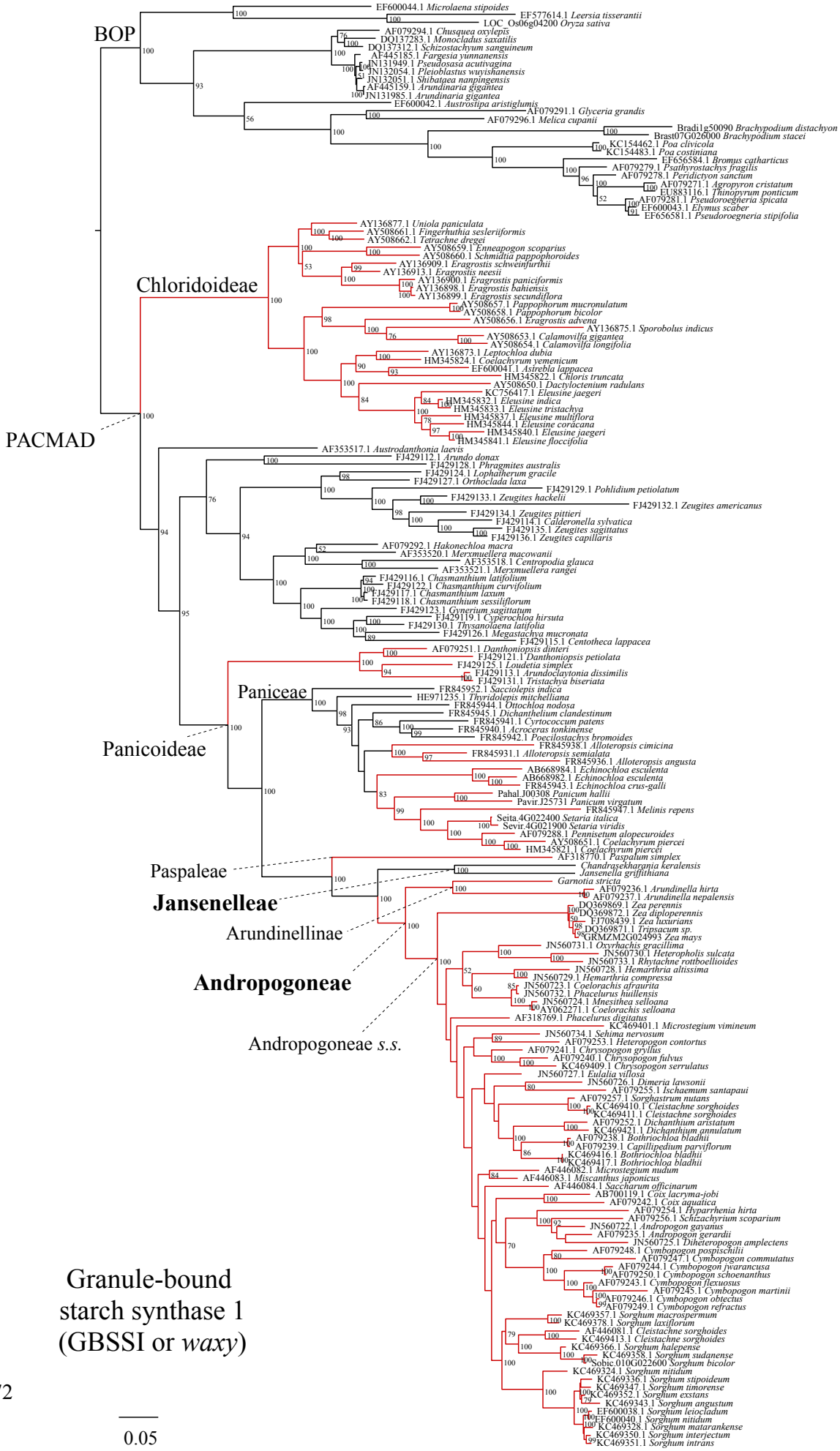

BOP
PACMAD
Paniceae
Paspaleae
Jansenelleae
Arundinellinae
Andropogoneae
Andropogoneae s.s.
LOC_Os06g35050 Oryza sativa
Bradi1g39160 Brachypodium distachyon
Brast07G132000 Brachypodium stacei
Sevir.4G143200 Setaria viridis
Seita.4G174900 Setaria italica
Pahal.D01470 Panicum hallii
Pavir.Db01059 Panicum virgatum
HE971215.1 Ancistrachne uncinulata
HE971216.1 Thyridolepis mitchelliana
HE971219.1 Neurachne annularis
FR845977.1 Ichnanthus vicinus
Jansenella griffithiana
Chandrasekharania keralensis
Garnotia stricta
GRMZM2G324297 Zea mays
Sobic.010G160400 Sorghum bicolor
FR845976.1 Hyparrhenia hirta
Arogenate dehydrogenase (arodeh)
0.05


BOP
PACMAD
Paniceae
Jansenelleae
Arundinellinae
Andropogoneae
Andropogoneae s.s.
DQ317422.1 Lithachne humilis
Bradi1g10047 Brachypodium distachyon
Brast02G291200 Brachypodium stacei
DQ317421.1 Chasmanthium latifolium
Pavir.Ia00733 Panicum virgatum
Pahal.I01913 Panicum hallii
DQ317419.1 Pennisetum glaucum
Seita.9G109500 Setaria italica
Sevir.9G107600 Setaria viridis
Jansenella griffithiana
Chandrasekharania keralensis
Garnotia stricta
GRMZM2G017087 Zea mays
Sobic.001G106200 Sorghum bicolor
KT449549.1 Eriochrysis holcoides
KT449568.1 Eriochrysis warmingiana
KT449545.1 Eriochrysis cayennensis
KT449567.1 Eriochrysis villosa
knotted 1 (kn1)
0.05


Brast09G192400 Brachypodium stacei
Bradi5g20340 Brachypodium distachyon
BOP
LOC_Os04g51000 Oryza sativa
KM264789.1 Bonia saxatilis
KM264736.1 Bambusa ventricosa
KM264759.1 Acidosasa purpurea
KM264748.1 Gaoligongshania megalothyrsa
KM264785.1 Ampelocalamus patellaris
KM264764.1 Arundinaria qingchengshanensis
KM264769.1 Chimonobambusa szechuanensis
KM264773.1 Yushania qiaojiaensis
KM264781.1 Pleioblastus maculatus
KF427096.1 Phyllostachys edulis
KF427095.1 Phyllostachys nigra
Paniceae
Sevir.7G234000 Setaria viridis
Seita.7G222300 Setaria italica
Pavir.Ga00621 Panicum virgatum
Pahal.G02162 Panicum hallii
AY789625.1 Arundinella hirta
PACMAD
Jansenella griffithiana
Chandrasekharania keralensis
Garnotia stricta
Jansenelleae
AY789596.1 Tripsacum dactyloides
AY789595.1 Tripsacum andersonii
GRMZM2G098813 Zea mays
AY789594.1 Zea luxurians
Arundinellinae
AY789592.1 Tripsacum latifolium
Andropogoneae
AY789590.1 Tripsacum floridanum
AY789591.1 Tripsacum zopilotense
AY789611.1 Chrysopogon fulvus
Andropogoneae s.s.
AY789606.1 Coelorachis aurita
AY789608.1 Coelorachis selloana
AY789604.1 Elionurus tripsacoides
AY789605.1 Elionurus muticus
AY789598.1 Zea diploperennis
GRMZM2G180190 Zea mays
AY789597.1 Zea luxurians
AY789600.1 Tripsacum andersonii
AY789601.1 Tripsacum floridanum
AY789602.1 Tripsacum zopilotense
AY789607.1 Apluda mutica
AY789613.1 Phacelurus digitatus
AY789614.1 Chionachne koenigii
AY789624.1 Coix lacryma-jobi
AY789609.1 Coix aquatica
AY789620.1 Ischaemum afrum
AY789622.1 Saccharum officinarum
Sobic.006G201600 Sorghum bicolor
AY789619.1 Cleistachne sorghoides
AY789621.1 Sorghum versicolor
AY789615.1 Hyparrhenia hirta
AY789617.1 Cymbopogon flexuosus
AY789610.1 Cymbopogon distans
AY789612.1 Heteropogon contortus
AY789616.1 Bothriochloa odorata
AY789618.1 Capillipedium parviflorum
floricaula/leafy-like
(floricaula)
0.05


BOP
PACMAD
Jansenelleae
Andropogoneae
retarded palea 1
(rep1)
0.05
LOC_Os09g24480 Oryza sativa
Bradi4g29980 Brachypodium distachyon
Brast05G121600 Brachypodium stacei
KM578622.1 Paspalum malacophyllum
Pahal.B02747 Panicum hallii
Sevir.2G210100 Setaria viridis
Seita.2G202000 Setaria italica
KM578624.1 Plagiantha tenella
Jansenella griffithiana
Chandrasekharania keralensis
KM578668.1 Arundinella hirta
Sobic.002G198400 Sorghum bicolor
Garnotia stricta
GRMZM2G064628 Zea mays
GRMZM2G110242 Zea mays
100
100
100
92
100
100
97
100
100
99
72
100
89


BOP
PACMAD
Paniceae
Paspaleae
Jansenelleae
Andropogoneae
Arundinellinae
Andropogoneae s.s.
LOC_Os03g49990 Oryza sativa
Brast02G278200 Brachypodium stacei
Bradi1g11090 Brachypodium distachyon
Sevir.9G121800 Setaria viridis
Seita.9G123000 Setaria italica
Pavir.Ia00840 Panicum virgatum
Pahal.I02073 Panicum hallii
KM578189.1 Plagiantha tenella
KM578185.1 Paspalum malacophyllum
Jansenella griffithiana
Chandrasekharania keralensis
KM578248.1 Arundinella hirta
KM578270.1 Arundinella nepalensis
Garnotia stricta
KM578258.1 Chrysopogon zizanioides
KM578184.1 Chrysopogon serrulatus
KM578290.1 Arthraxon lanceolatus
KM578257.1 Arthraxon prionodes
JQ951708.1 Mnesithea lepidura
KM578145.1 Phacelurus huillensis
GRMZM2G144744 Zea mays
KM578154.1 Tripsacum dactyloides
GRMZM2G024973 Zea mays
KP233124.1 Polytoca wallichiana
KP233123.1 Chionachne koenigii
KM578292.1 Chionachne massiei
KM578156.1 Rottboellia cochinchinensis
KM578275.1 Chasmopodium caudatum
KM578175.1 Coix lacryma-jobi
KM578293.1 Coix gasteenii
KM578251.1 Apluda mutica
KM578142.1 Ischaemum afrum
KM578168.1 Andropterum stolzii
KM578323.1 Ischaemum indicum
KM578356.1 Ischaemum rugosum
KM578143.1 Ischaemum santapaui
KM578302.1 Dimeria fuscescens
KM578304.1 Dimeria ornithopoda
KM578306.1 Dimeria sinensis
KT449489.1 Eriochrysis warmingiana
KT449465.1 Eriochrysis holcoides
KT449481.1 Eriochrysis cayennensis
KT449488.1 Eriochrysis villosa
KT449474.1 Eriochrysis laxa
KM578137.1 Eriochrysis pallida
KM578269.1 Saccharum ravennae
KM578333.1 Saccharum arundinaceum
KM578328.1 Kerriochloa siamensis
KM578208.1 Polytrias indica
KM578206.1 Polytrias indica
KM578247.1 Eulalia aurea
KM578321.1 Imperata cylindrica
KT449491.1 Imperata tenuis
KT449490.1 Imperata brasiliensis
KM578342.1 Spodiopogon sibiricus
KM578203.1 Pogonatherum paniceum
KM578289.1 Apocopis siamensis
KM578282.1 Apocopis courtallumensis
KM578286.1 Apocopis intermedius
KM578312.1 Germainia capitata
KM578284.1 Apocopis intermedius
KM578285.1 Apocopis intermedius
KM578338.1 Saccharum narenga
KM578331.1 Pseudosorghum fasciculare
KM578200.1 Miscanthus sinensis
Sobic.001G120900 Sorghum bicolor
KR493982.1 Sorghum laxiflorum
KP233106.1 Saccharum asperum
KP233110.1 Saccharum asperum
KP233112.1 Saccharum giganteum
KP233121.1 Saccharum villosum
KP233100.1 Saccharum angustifolium
KM578230.1 Miscanthus ecklonii
KM578119.1 Andropogon eucomus
KM578362.1 Andropogon glomeratus
KM578209.1 Andropogon virginicus
KM578170.1 Andropogon gayanus
KM578121.1 Andropogon perligulatus
KM578277.1 Andropogon ascinodis
JQ951714.1 Phacelurus digitatus
KM578124.1 Schizachyrium sanguineum
KM578155.1 Schizachyrium exile
KM578172.1 Andropogon hallii
KM578120.1 Andropogon gerardii
KM578182.1 Diheteropogon amplectens
KM578355.1 Diheteropogon hagerupii
KM578318.1 Hyparrhenia diplandra
KM578359.1 Hyparrhenia hirta
KM578319.1 Hyparrhenia rufa
KR493970.1 Sorghum plumosum
KR493972.1 Sorghum plumosum
KR493971.1 Sorghum plumosum
KR493984.1 Sorghum brachypodum
KR493983.1 Sorghum brachypodum
KR493978.1 Sorghum matarankense
KR493968.1 Sorghum stipoideum
KR493974.1 Sorghum exstans
KR493977.1 Sorghum matarankense
KR493976.1 Sorghum angustum
KM578161.1 Eulalia villosa
KM578311.1 Eulalia quadrinervis
KM578291.1 Capillipedium assimile
KM578357.1 Dichanthium annulatum
KM578298.1 Dichanthium theinlwinii
KM578136.1 Dichanthium aristatum
KM578129.1 Cymbopogon flexuosus
KM578134.1 Cymbopogon refractus
KM578130.1 Cymbopogon obtectus
KM578131.1 Cymbopogon obtectus
KM578163.1 Cymbopogon martinii
KM578296.1 Cymbopogon citratus
KM578294.1 Cymbopogon citratus
KM578126.1 Cymbopogon distans
KM578167.1 Cymbopogon jwarancusa
KM578132.1 Cymbopogon parkeri
KM578148.1 Sehima nervosum
KM578315.1 Heteropogon triticeus
KM578346.1 Themeda arundinacea
KM578272.1 Themeda triandra
KM578193.1 Iseilema macratherum
KM578351.1 Dichanthium foveolatum
KM578225.1 Bothriochloa laguroides
KM578220.1 Bothriochloa woodrovii
KM578235.1 Bothriochloa hybrida
KM578211.1 Bothriochloa alta
KM578244.1 Bothriochloa exaristata
KM578212.1 Bothriochloa barbinodis
KM578197.1 Bothriochloa ischaemum
KM578213.1 Bothriochloa macra
KM578228.1 Bothriochloa decipiens
KM578216.1 Bothriochloa insculpta
KM578237.1 Bothriochloa bladhii
KM578123.1 Capillipedium spicigerum
JQ951709.1 Chrysopogon gryllus
DELLA protein-encoding gene dwarf 8 (dwarf8)
0.05


BOP
Chloridoideae
PACMAD
Paniceae
Panicoideae
Paspaleae
Jansenelleae
Arundinellinae
Andropogoneae
Andropogoneae s.s.
EF600044.1 Microlaena stipoides
EF577614.1 Leersia tisserantii
LOC_Os06g04200 Oryza sativa
AF079294.1 Chusquea oxylepis
DQ137283.1 Monocladus saxatilis
DQ137312.1 Schizostachyum sanguineum
AF445185.1 Fargesia yunnanensis
JN131949.1 Pseudoxasa acutivagina
JN132054.1 Pleioblastus wuyishanensis
JN132051.1 Shibataea nanpingensis
AF445159.1 Arundinaria gigantea
JN131985.1 Arundinaria gigantea
EF600042.1 Austrostipa aristiglumis
AF079291.1 Glyceria grandis
AF079296.1 Melica cupanii
Bradi1g50090 Brachypodium distachyon
Brast07G026000 Brachypodium stacei
KC154462.1 Poa clivicola
KC154483.1 Poa costiniana
EF656584.1 Bromus catharticus
AF079279.1 Psathyrostachys fragilis
AF079278.1 Peridictyon sanctum
AF079271.1 Agropyron cristatum
EU883116.1 Thinopyrum ponticum
AF079281.1 Pseudoroegneria spicata
EF600043.1 Elymus scaber
EF656581.1 Pseudoroegneria stipifolia
AY136877.1 Uniola paniculata
AY508661.1 Fingerhuthia sesleriiformis
AY508662.1 Tetrachne dregei
AY508659.1 Enneapogon scoparius
AY508660.1 Schmidtia pappophoroides
AY136909.1 Eragrostis schweinfurthii
AY136913.1 Eragrostis neesii
AY136900.1 Eragrostis paniciformis
AY136898.1 Eragrostis bahiensis
AY136899.1 Eragrostis secundiflora
AY508657.1 Pappophorum mucronulatum
AY508658.1 Pappophorum bicolor
AY508656.1 Eragrostis advena
AY136875.1 Sporobolus indicus
AY508653.1 Calamovilfa gigantea
AY508654.1 Calamovilfa longifolia
AY136873.1 Leptochloa dubia
HM345824.1 Coelachyrum yemenicum
EF600041.1 Astrebla lappacea
HM345822.1 Chloris truncata
AY508650.1 Dactyloctenium radulans
KC756417.1 Eleusine jaegeri
HM345832.1 Eleusine indica
HM345833.1 Eleusine tristachya
HM345837.1 Eleusine multiflora
HM345844.1 Eleusine coracana
HM345840.1 Eleusine jaegeri
HM345841.1 Eleusine floccifolia
AF353517.1 Austrodanthonia laevis
FJ429112.1 Arundo donax
FJ429128.1 Phragmites australis
FJ429124.1 Lophatherum gracile
FJ429127.1 Orthoclada laxa
FJ429129.1 Pohlidium petiolatum
FJ429133.1 Zeugites hackelii
FJ429132.1 Zeugites americanus
FJ429134.1 Zeugites pittieri
FJ429114.1 Calderonella sylvatica
FJ429135.1 Zeugites sagittatus
FJ429136.1 Zeugites capillaris
AF079292.1 Hakonechloa macra
AF353520.1 Merxmuellera macowanii
AF353518.1 Centropodia glauca
AF353521.1 Merxmuellera rangei
FJ429116.1 Chasmanthium latifolium
FJ429122.1 Chasmanthium curvifolium
FJ429117.1 Chasmanthium laxum
FJ429118.1 Chasmanthium sessiliflorum
FJ429123.1 Gynerium sagittatum
FJ429119.1 Cyperochloa hirsuta
FJ429130.1 Thysanolaena latifolia
FJ429126.1 Megastachya mucronata
FJ429115.1 Centotheca lappacea
AF079251.1 Danthoniopsis dinteri
FJ429121.1 Danthoniopsis petiolata
FJ429125.1 Loudetia simplex
FJ429113.1 Arundoclaytonia dissimilis
FJ429131.1 Tristachya biseriata
FR845952.1 Sacciolepis indica
HE971235.1 Thyridolepis mitchelliana
FR845944.1 Ottochloa nodosa
FR845945.1 Dichanthelium clandestinum
FR845941.1 Cyrtococcum patens
FR845940.1 Acroceras tonkinense
FR845942.1 Poecilostachys bromoides
FR845938.1 Alloteropsis cimicina
FR845931.1 Alloteropsis semialata
FR845936.1 Alloteropsis angusta
AB668984.1 Echinochloa esculenta
AB668982.1 Echinochloa esculenta
FR845943.1 Echinochloa crus-galli
Pahal.J00308.Panicum hallii
Pavir.J25731.Panicum virgatum
FR845947.1 Melinis repens
Seita.4G022400.Setaria italica
Sevir.4G021900.Setaria viridis
AF079288.1 Pennisetum alopecuroides
AY508651.1 Coelachyrum piercei
HM345821.1 Coelachyrum piercei
AF318770.1 Paspalum simplex
Chandrasekharania keralensis
Jansenella griffithiana
Garnotia stricta
AF079236.1 Arundinella hirta
AF079237.1 Arundinella nepalensis
DQ369869.1 Zea perennis
DQ369872.1 Zea diploperennis
FJ708439.1 Zea luxurians
DQ369871.1 Tripsacum sp.
GRMZM2G024993 Zea mays
JN560731.1 Oxyrhachis gracillima
JN560730.1 Heteropholis sulcata
JN560733.1 Rhytachne rottboellioides
JN560728.1 Hemarthria altissima
JN560729.1 Hemarthria compressa
JN560723.1 Coelorachis afraurita
JN560732.1 Phacelurus huillensis
JN560724.1 Mnesithea selloana
AY062271.1 Coelorachis selloana
AF318769.1 Phacelurus digitatus
KC469401.1 Microstegium vimineum
JN560734.1 Sehima nervosum
AF079253.1 Heteropogon contortus
AF079241.1 Chrysopogon gryllus
AF079240.1 Chrysopogon fulvus
KC469409.1 Chrysopogon serrulatus
JN560727.1 Eulalia villosa
JN560726.1 Dimeria lawsonii
AF079255.1 Ischaemum santapaui
AF079257.1 Sorghastrum nutans
KC469410.1 Cleistachne sorghoides
KC469411.1 Cleistachne sorghoides
AF079252.1 Dichanthium aristatum
KC469421.1 Dichanthium annulatum
AF079238.1 Bothriochloa bladhii
AF079239.1 Capillipedium parviflorum
KC469416.1 Bothriochloa bladhii
KC469417.1 Bothriochloa bladhii
AF446082.1 Microstegium nudum
AF446083.1 Miscanthus japonicus
AF446084.1 Saccharum officinarum
AB700119.1 Coix lacryma-jobi
AF079242.1 Coix aquatica
AF079254.1 Hyparrhenia hirta
AF079256.1 Schizachyrium scoparium
JN560722.1 Andropogon gayanus
AF079235.1 Andropogon gerardii
JN560725.1 Diheteropogon amplectens
AF079248.1 Cymbopogon pospischilii
AF079247.1 Cymbopogon commutatus
AF079244.1 Cymbopogon jwarancusa
AF079250.1 Cymbopogon schoenanthus
AF079243.1 Cymbopogon flexuosus
AF079245.1 Cymbopogon martinii
AF079246.1 Cymbopogon obtectus
AF079249.1 Cymbopogon refractus
KC469357.1 Sorghum macrospermum
KC469378.1 Sorghum laxiflorum
AF446081.1 Cleistachne sorghoides
KC469413.1 Cleistachne sorghoides
KC469366.1 Sorghum halepense
KC469358.1 Sorghum sudanense
Sobic.010G022600 Sorghum bicolor
KC469324.1 Sorghum nitidum
KC469336.1 Sorghum stipoideum
KC469347.1 Sorghum timorense
KC469352.1 Sorghum exstans
KC469343.1 Sorghum angustum
EF600038.1 Sorghum leiocladum
EF600040.1 Sorghum nitidum
KC469328.1 Sorghum matarankense
KC469350.1 Sorghum interjectum
KC469351.1 Sorghum intrans
Granule-bound starch synthase 1 (GBSSI or waxy)
0.05

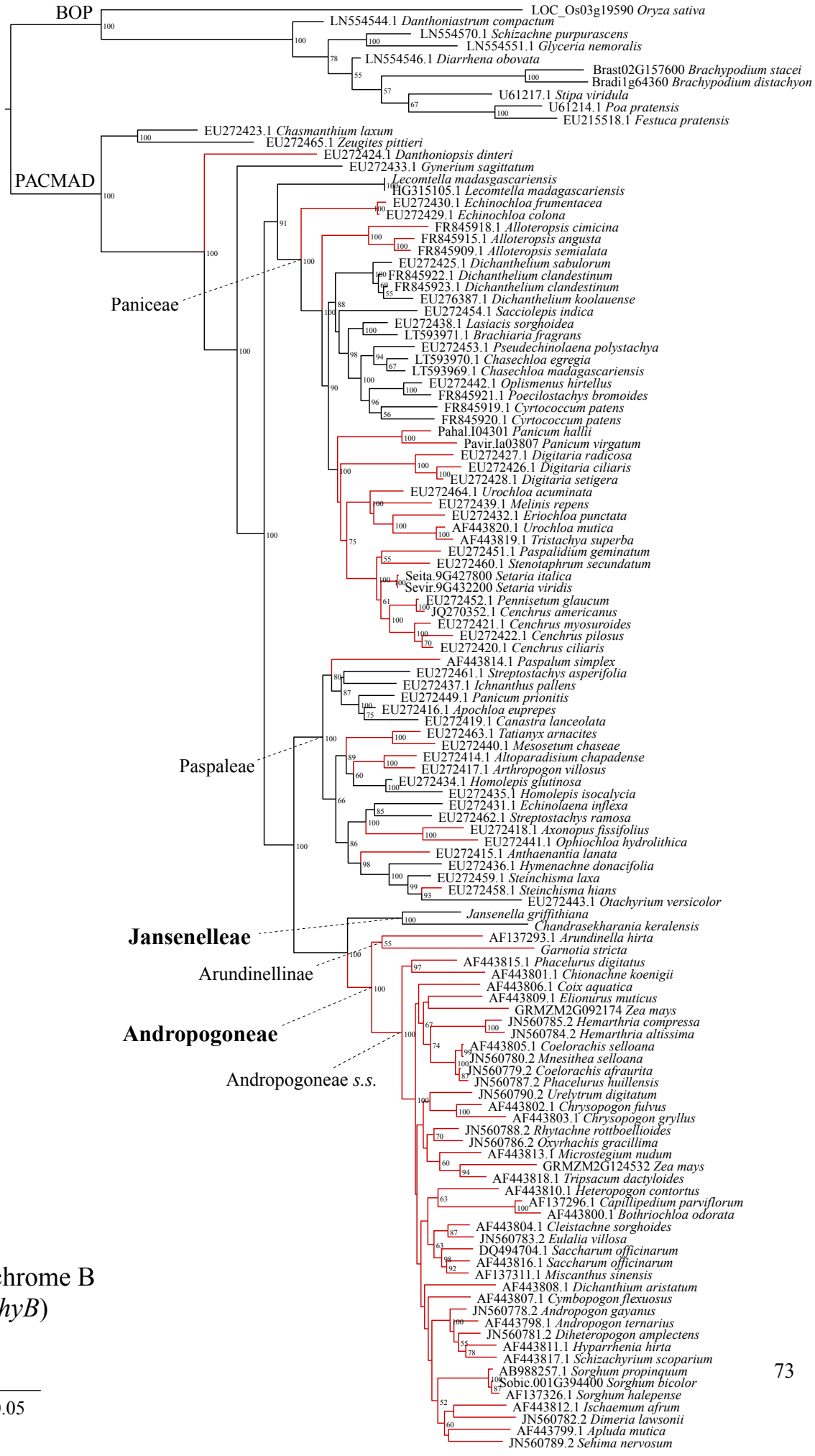

BOP
PACMAD
Paniceae
Paspaleae
Jansenelleae
Arundinellinae
Andropogoneae
Andropogoneae s.s.
LOC_Os03g19590 Oryza sativa
LN554544.1 Danthoniastrum compactum
LN554570.1 Schizachne purpurascens
LN554551.1 Glyceria nemoralis
LN554546.1 Diarrhena obovata
Brast02G157600 Brachypodium stacei
Bradi1g64360 Brachypodium distachyon
U61217.1 Stipa viridula
U61214.1 Poa pratensis
EU215518.1 Festuca pratensis
EU272423.1 Chasmanthium laxum
EU272465.1 Zeugites pittieri
EU272424.1 Danthoniopsis dinteri
EU272433.1 Gynerium sagittatum
Lecomtella madasgascariensis
HG315105.1 Lecomtella madagascariensis
EU272430.1 Echinochloa frumentacea
EU272429.1 Echinochloa colona
FR845918.1 Alloteropsis cimicina
FR845915.1 Alloteropsis angusta
FR845909.1 Alloteropsis semialata
EU272425.1 Dichanthelium sabulorum
FR845922.1 Dichanthelium clandestinum
FR845923.1 Dichanthelium clandestinum
EU276387.1 Dichanthelium koolauense
EU272454.1 Sacciolepis indica
EU272438.1 Lasiacis sorghoidea
LT593971.1 Brachiaria fragrans
EU272453.1 Pseudechinolaena polystachya
LT593970.1 Chasechloa egregia
LT593969.1 Chasechloa madagascariensis
EU272442.1 Oplismenus hirtellus
FR845921.1 Poecilostachys bromoides
FR845919.1 Cyrtococcum patens
FR845920.1 Cyrtococcum patens
Pahal.I04301 Panicum hallii
Pavir.Ia03807 Panicum virgatum
EU272427.1 Digitaria radicosa
EU272426.1 Digitaria ciliaris
EU272428.1 Digitaria setigera
EU272464.1 Urochloa acuminata
EU272439.1 Melinis repens
EU272432.1 Eriochloa punctata
AF443820.1 Urochloa mutica
AF443819.1 Tristachya superba
EU272451.1 Paspalidium geminatum
EU272460.1 Stenotaphrum secundatum
Seita.9G427800 Setaria italica
Sevir.9G432200 Setaria viridis
EU272452.1 Pennisetum glaucum
JQ270352.1 Cenchrus americanus
EU272421.1 Cenchrus myosuroides
EU272422.1 Cenchrus pilosus
EU272420.1 Cenchrus ciliaris
AF443814.1 Paspalum simplex
EU272461.1 Streptostachys asperifolia
EU272437.1 Ichnanthus pallens
EU272449.1 Panicum prionitis
EU272416.1 Apochloa euprepes
EU272419.1 Canastra lanceolata
EU272463.1 Tatianyx arnacites
EU272440.1 Mesosetum chaseae
EU272414.1 Altoparadisium chapadense
EU272417.1 Arthropogon villosus
EU272434.1 Homolepis glutinosa
EU272435.1 Homolepis isocalycia
EU272431.1 Echinolaena inflexa
EU272462.1 Streptostachys ramosa
EU272418.1 Axonopus fissifolius
EU272441.1 Ophiochloa hydrolithica
EU272415.1 Anthaenantia lanata
EU272436.1 Hymenachne donacifolia
EU272459.1 Steinchisma laxa
EU272458.1 Steinchisma hians
EU272443.1 Otachyrium versicolor
Jansenella griffithiana
Chandrasekharania keralensis
AF137293.1 Arundinella hirta
Garnotia stricta
AF443815.1 Phacelurus digitatus
AF443801.1 Chionachne koenigii
AF443806.1 Coix aquatica
AF443809.1 Elionurus muticus
GRMZM2G092174 Zea mays
JN560785.2 Hemarthria compressa
JN560784.2 Hemarthria altissima
AF443805.1 Coelorachis selloana
JN560780.2 Mnesithea selloana
JN560779.2 Coelorachis afraurita
JN560787.2 Phacelurus huillensis
JN560790.2 Urelytrum digitatum
AF443802.1 Chrysopogon fulvus
AF443803.1 Chrysopogon gryllus
JN560788.2 Rhytachne rottboellioides
JN560786.2 Oxyrhachis gracillima
AF443813.1 Microstegium nudum
GRMZM2G124532 Zea mays
AF443818.1 Tripsacum dactyloides
AF443810.1 Heteropogon contortus
AF137296.1 Capillipedium parviflorum
AF443800.1 Bothriochloa odorata
AF443804.1 Cleistachne sorghoides
JN560783.2 Eulalia villosa
DQ494704.1 Saccharum officinarum
AF443816.1 Saccharum officinarum
AF137311.1 Miscanthus sinensis
AF443808.1 Dichanthium aristatum
AF443807.1 Cymbopogon flexuosus
JN560778.2 Andropogon gayanus
AF443798.1 Andropogon ternarius
JN560781.2 Diheteropogon amplectens
AF443811.1 Hyparrhenia hirta
AF443817.1 Schizachyrium scoparium
AB988257.1 Sorghum propinquum
Sobic.001G394400 Sorghum bicolor
AF137326.1 Sorghum halepense
AF443812.1 Ischaemum afrum
JN560782.2 Dimeria lawsonii
AF443799.1 Apluda mutica
JN560789.2 Sehima nervosum
Phytochrome B
(phyB)
0.05

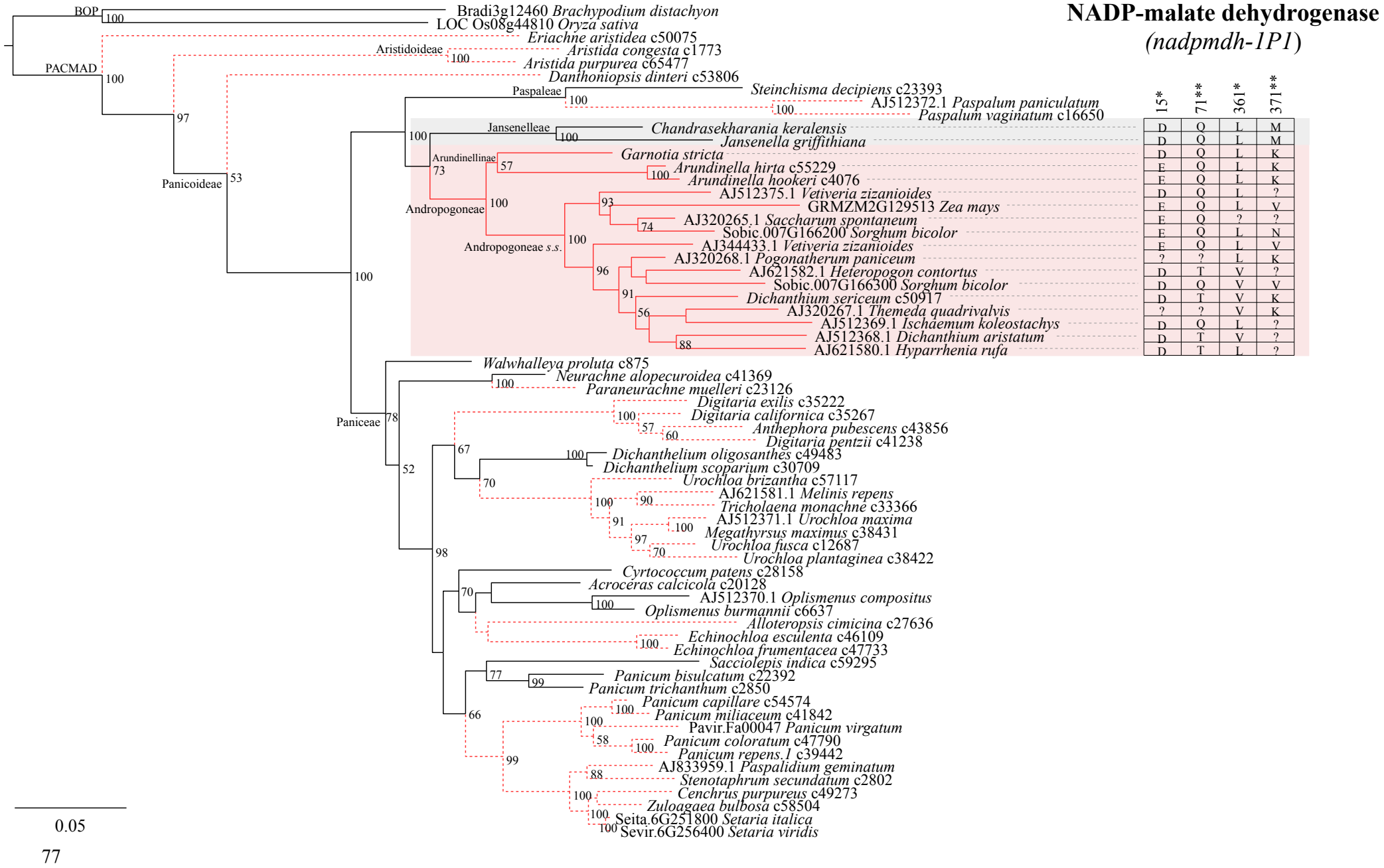
Fig. 2.S4. Bayesian trees of genes encoding C_4 enzymes inferred from 3rd positions of codons. Amino acid sites under positive selection according to a Bayes Empirical Bayes test are show on the right, with asterisks indicating posterior probabilities (* > 95% and ** > 99%). Red and black branches are C_4 and C_3 species, respectively. Dashed branches are C_4 species outside Andropogoneae that were pruned before positive selection analyses. Posterior probabilities are shown near nodes (values < 50% were omitted).

NADP-malic enzyme
(nadpme-1P4)

5*	6**	10**	49**	50*	60**	61**	67*	95*	100**	118**	149**	184*	186*	201*	222*	254**	257**	258*	260*	295**	317**	350**	364*	398*	439**	446*	484**	530*
V	T	F	V	S	H	N	V	F	V	R	V	R	S	E	H	D	A	K	S	H	I	Q	T	S	D	K	E	R
V	T	F	V	S	H	N	V	F	V	R	V	R	S	E	H	D	A	K	S	H	I	Q	T	S	D	K	E	R
I	T	F	V	P	H	N	?	F	V	R	V	R	S	E	H	D	A	N	S	H	I	Q	T	S	N	K	A	R
V	T	S	I	S	Y	N	V	F	Y	R	C	?	?	K	D	D	E	T	S	Q	M	Q	K	S	A	K	E	R
?	?	?	V	S	H	N	V	F	?	?	V	?	?	D	H	D	S	K	S	?	?	Q	T	S	?	N	?	?
I	T	F	V	S	N	S	V	F	V	R	C	S	L	Q	D	E	E	K	S	Q	M	Q	T	S	K	T	E	R
I	T	F	V	S	N	S	V	F	V	R	C	S	L	Q	D	E	E	K	S	Q	M	Q	T	S	K	T	E	R
?	?	?	?	?	H	?	?	?	?	R	C	S	L	K	D	D	E	K	S	?	?	?	S	P	N	?	?	?
?	?	?	V	S	N	S	V	F	V	?	C	?	?	D	E	E	R	S	?	?	Q	T	S	P	T	?	?	?
?	?	?	?	?	?	?	?	F	V	R	C	R	S	K	D	D	A	K	S	Q	I	Q	S	S	A	K	?	?
I	L	F	V	P	H	N	V	F	V	R	C	R	S	K	H	N	A	K	S	H	I	Q	T	S	D	K	E	R
?	?	?	?	?	?	?	?	F	V	R	C	R	S	K	H	D	A	K	S	Q	I	Q	T	S	D	K	?	?
I	M	F	V	P	H	N	V	F	V	R	C	R	S	K	H	D	A	K	S	H	I	Q	T	S	D	K	E	R
?	?	?	?	?	?	?	?	F	V	R	C	R	S	K	H	D	A	K	R	Q	I	Q	T	S	D	K	?	?
?	?	?	?	?	?	?	?	F	I	R	C	R	S	N	H	D	A	K	R	H	I	Q	T	S	D	K	?	?
?	?	?	?	?	?	?	?	?	V	V	R	C	R	S	K	D	D	E	K	S	Q	I	Q	T	S	D	K	?	?
V	M	T	L	S	N	T	T	V	F	G	C	D	S	K	D	D	E	K	S	Q	I	Q	K	S	A	K	K	R
?	?	?	?	?	?	?	?	F	V	R	C	R	S	K	D	D	E	K	S	Q	I	A	G	S	D	K	?	?
?	?	?	?	?	?	?	?	F	V	R	C	R	S	K	D	D	E	K	R	Q	I	A	G	S	D	K	?	?
V	T	F	V	S	H	N	V	F	V	R	C	R	S	K	D	D	E	K	R	Q	I	Q	G	S	A	K	A	R
?	?	?	?	?	?	?	?	F	G	C	D	S	K	D	D	E	K	S	Q	I	A	T	T	A	K	?	?	
?	?	?	?	?	?	?	?	V	Y	G	C	D	T	K	D	D	E	K	S	Q	M	A	T	S	A	K	?	?
V	M	T	V	S	N	N	L	V	Y	G	C	D	S	K	D	D	E	K	S	Q	M	A	T	S	A	K	A	R
?	?	?	?	?	?	?	?	V	F	G	C	D	S	K	D	E	D	K	S	Q	M	A	T	T	A	K	?	?
V	T	S	A	T	N	I	L	V	Y	S	C	D	T	K	D	E	E	K	S	Q	M	A	T	A	A	K	T	K
?	?	?	?	?	?	?	?	V	F	G	C	D	T	K	D	E	E	K	S	Q	V	A	T	T	P	K	?	?

Phosphoenolpyruvate carboxylase
(ppc-1P3)

20**	27**	29**	31**	32**	35**	49**	61**	98**	115**	122**	161**	172**	222**	226**	227**	232**	257**	342**	351*	393**	401**	404**	427**	434**
E	Q	W	L	S	R	Q	R	R	A	R	K	R	S	E	H	T	A	H	I	E	Q	N	P	R
E	Q	W	L	S	R	Q	R	R	A	R	K	R	S	E	H	T	A	H	I	D	Q	N	P	R
S	Q	W	L	S	R	Q	Q	R	A	R	R	R	S	E	N	T	A	Q	F	D	Q	I	P	R
S	M	W	A	S	R	M	R	K	N	R	K	R	S	E	H	T	T	H	I	E	N	K	P	R
S	Q	W	V	S	R	M	R	R	N	R	K	K	S	E	H	T	T	H	I	E	N	K	P	R
E	Q	W	L	S	R	Q	H	R	S	R	R	C	E	H	T	K	F	F	E	Q	D	P	V	
?	?	F	V	A	Q	M	R	R	G	K	V	R	M	D	N	S	K	?	?	?	?	?	?	?
D	M	W	V	S	Q	Q	R	R	A	K	V	R	M	E	H	S	K	Y	F	D	Q	D	P	R
S	V	W	V	G	K	M	H	K	N	K	V	K	M	E	N	S	K	W	F	D	S	D	L	R
S	M	W	V	S	K	M	R	R	A	K	V	K	M	E	N	S	K	W	F	E	Q	D	P	R
S	M	W	I	S	K	M	K	K	N	R	T	R	C	E	H	T	A	A	N	D	E	K	P	R
S	F	W	L	S	R	M	R	R	G	K	V	K	C	E	T	T	A	F	S	?	?	?	?	?
S	M	W	V	S	K	M	H	R	A	R	V	K	M	D	H	S	K	W	F	D	Q	D	P	R
S	M	W	V	S	K	Q	H	K	A	R	V	R	M	D	N	S	K	W	F	E	Q	D	P	R
S	V	W	V	S	R	M	R	R	A	K	V	K	M	D	H	S	K	W	F	D	Q	E	P	R
S	V	W	V	M	R	M	R	R	A	C	V	K	M	D	N	S	R	W	F	D	Q	D	V	R
S	V	F	E	S	K	M	R	K	A	K	V	K	M	E	N	S	K	W	F	D	S	D	P	R
S	V	F	V	T	K	M	K	K	G	K	V	K	M	E	N	S	K	W	F	E	S	D	P	K
S	V	W	V	S	S	M	K	K	A	K	V	K	M	D	N	S	K	W	F	E	Q	E	P	R
S	V	F	V	S	K	M	K	K	G	K	V	K	M	D	N	S	K	W	F	D	S	D	P	R
S	V	F	V	S	K	M	K	K	G	K	V	K	M	E	N	S	K	W	F	E	S	D	P	R
S	V	F	V	S	K	M	K	K	G	K	V	K	M	D	N	S	K	W	F	D	S	D	P	R
S	V	F	V	S	K	M	K	K	G	K	V	K	M	E	N	S	K	W	F	D	S	D	P	K


NADP-malate dehydrogenase
(nadpmdh-1P1)
15*
71**
361*
371**
BOP
PACMAD
Panicoideae
Aristidoideae
Paspaleae
Jansenelleae
Arundinellinae
Andropogoneae
Andropogoneae s.s.
Paniceae
Bradi3g12460 Brachypodium distachyon
LOC Os08g44810 Oryza sativa
Eriachne aristidea c50075
Aristida congesta c1773
Aristida purpurea c65477
Danthoniopsis dinteri c53806
Steinchisma decipiens c23393
AJ512372.1 Paspalum paniculatum
Paspalum vaginatum c16650
Chandrasekharania keralensis
Jansenella griffithiana
Garnotia stricta
Arundinella hirta c55229
Arundinella hookeri c4076
AJ512375.1 Vetiveria zizanioides
GRMZM2G129513 Zea mays
AJ320265.1 Saccharum spontaneum
Sobic.007G166200 Sorghum bicolor
AJ344433.1 Vetiveria zizanioides
AJ320268.1 Pogonatherum paniceum
AJ621582.1 Heteropogon contortus
Sobic.007G166300 Sorghum bicolor
Dichanthium sericeum c50917
AJ320267.1 Themeda quadrivalvis
AJ512369.1 Ischaemum koleostachys
AJ512368.1 Dichanthium aristatum
AJ621580.1 Hyparrhenia rufa
Walwhalleya proluta c875
Neurachne alopecuroidea c41369
Paraneurachne muelleri c23126
Digitaria exilis c35222
Digitaria californica c35267
Anthephora pubescens c43856
Digitaria pentzii c41238
Dichanthelium oligosanthes c49483
Dichanthelium scoparium c30709
Urochloa brizantha c57117
AJ621581.1 Melinis repens
Tricholaena monachne c33366
AJ512371.1 Urochloa maxima
Megathyrsus maximus c38431
Urochloa fusca c12687
Urochloa plantaginea c38422
Cyrtococcum patens c28158
Acroceras calcicola c20128
AJ512370.1 Oplismenus compositus
Oplismenus burmannii c6637
Alloteropsis cimicina c27636
Echinochloa esculenta c46109
Echinochloa frumentacea c47733
Sacciolepis indica c59295
Panicum bisulcatum c22392
Panicum trichanthum c2850
Panicum capillare c54574
Panicum miliaceum c41842
Pavir.Fa00047 Panicum virgatum
Panicum coloratum c47790
Panicum repens.1 c39442
AJ833959.1 Paspalidium geminatum
Stenotaphrum secundatum c2802
Cenchrus purpureus c49273
Zuloagaea bulbosa c58504
Seita.6G251800 Setaria italica
Sevir.6G256400 Setaria viridis
0.05
77

Phosphoenolpyruvate carboxykinase
(pck-1P1)

Pyruvate, phosphate dikinase
(ppdk-1P2)

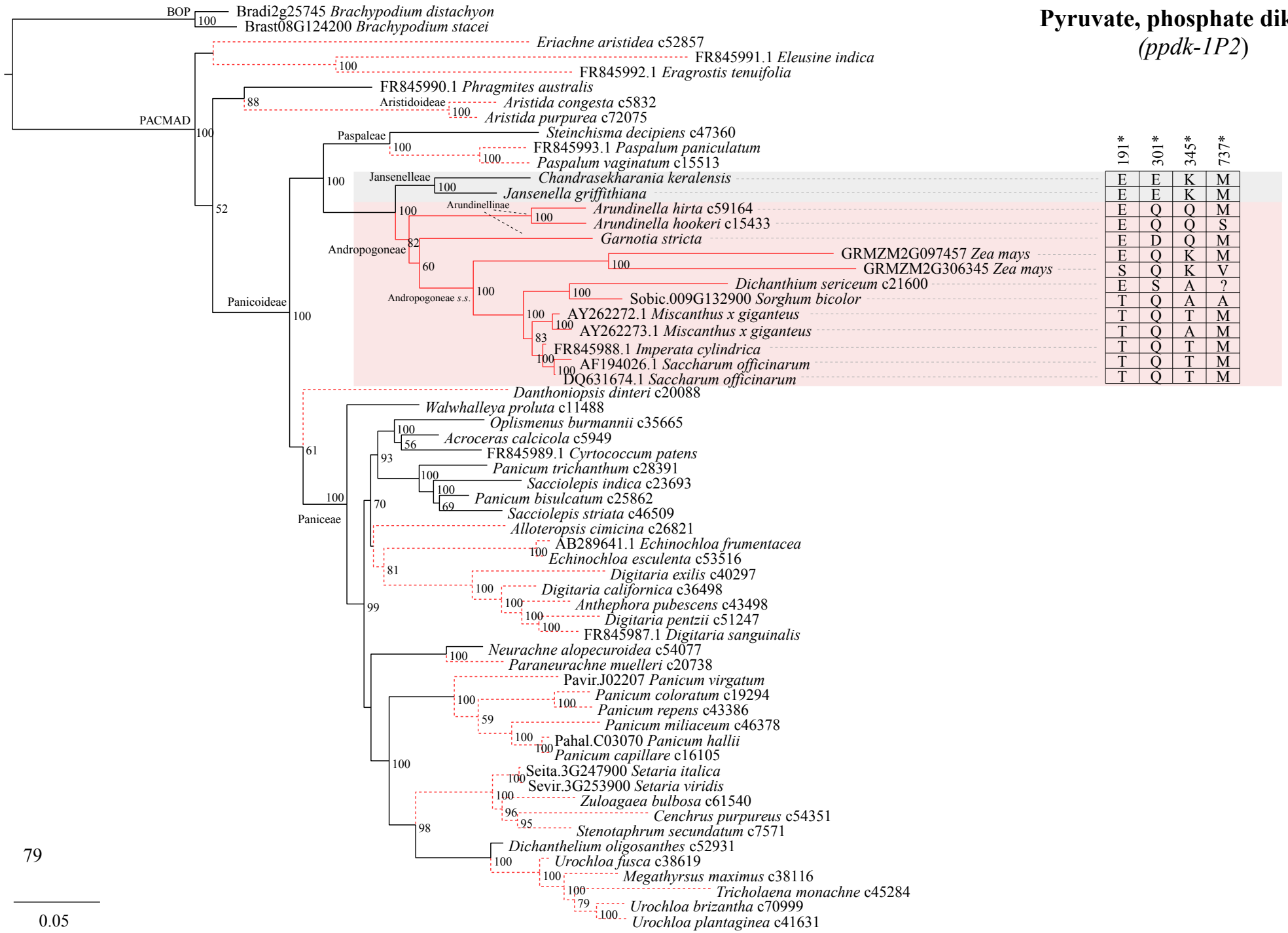
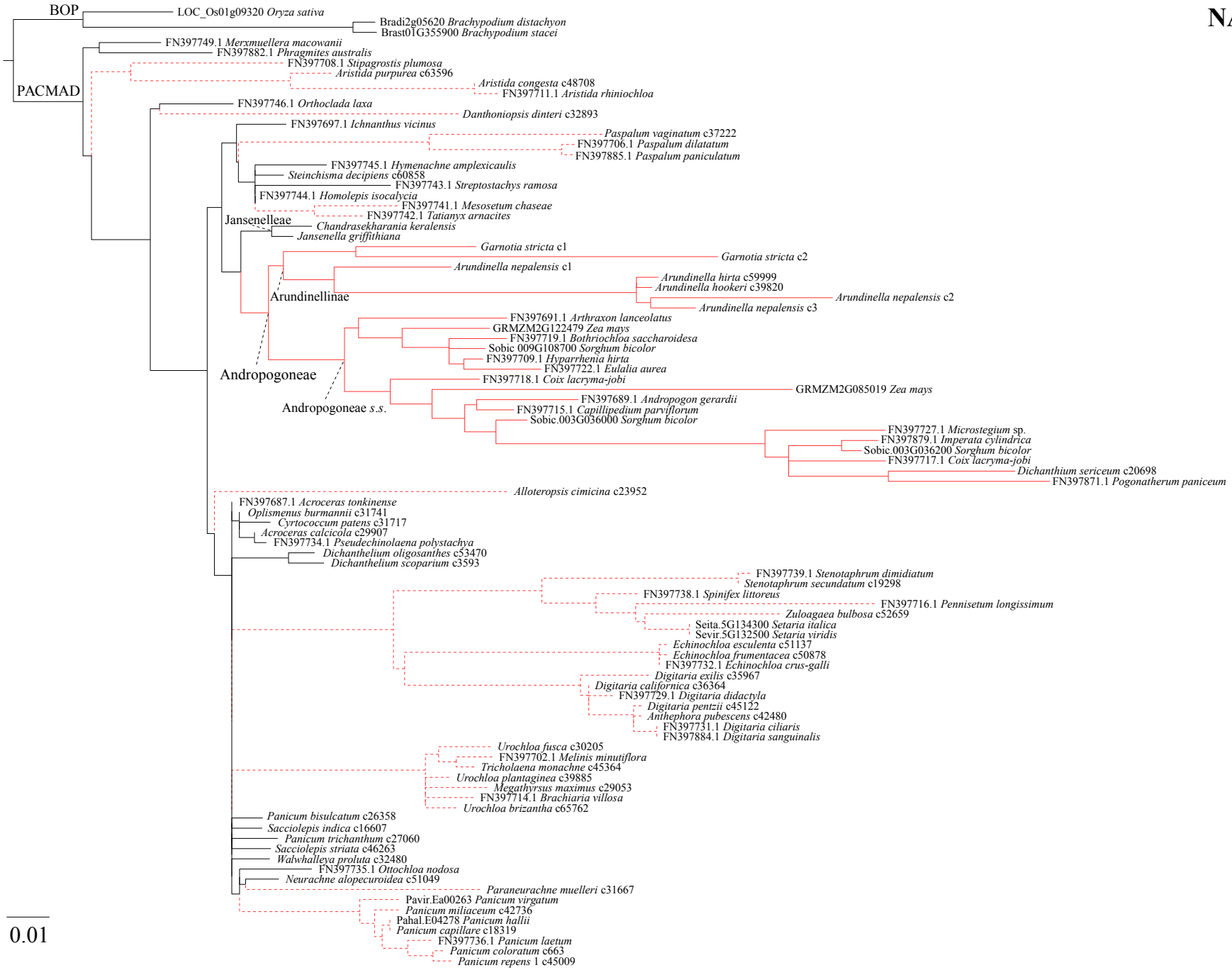


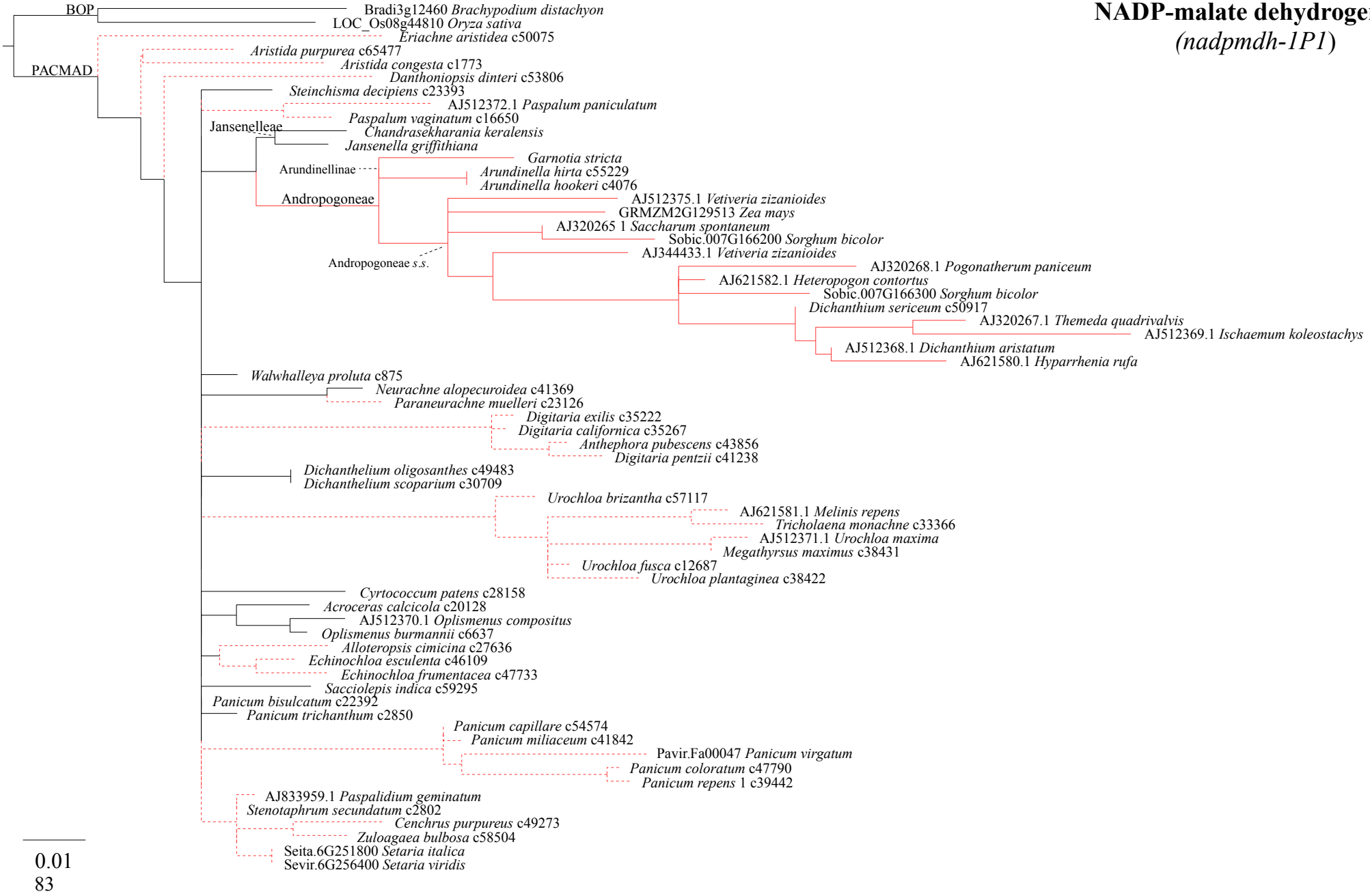
Fig. 2.S5. Protein trees with topologies constrained to that obtained using 3rd positions of codons for genes encoding five core C_4 enzymes: NADP-malate dehydrogenase (NADP-MDH), NADP-malic enzyme (NADP-ME), phosphoenolpyruvate carboxykinase (PCK), phosphoenolpyruvate carboxylase (PEPC) and pyruvate phosphate dikinase (PPDK). Scale bar = 0.01 amino acid substitutions per site. Branches are coloured in red for C_4 and black for C_3 accessions. Dashed branches are C_4 species outside Andropogoneae that were pruned before positive selection analyses.

Pyruvate, phosphate dikinase
(ppdk-1P2)

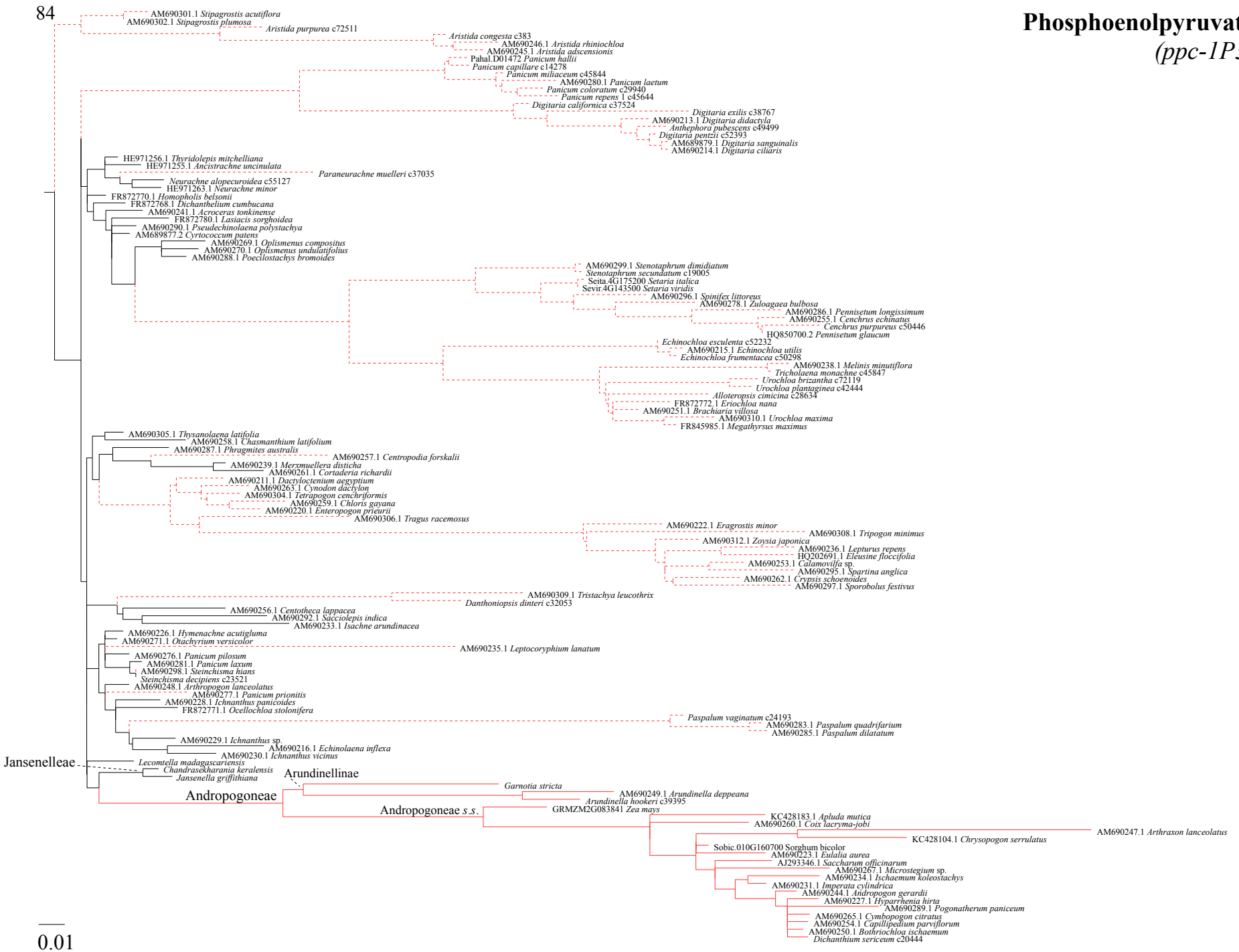
NADP-malic enzyme
(nadpme-1P4)



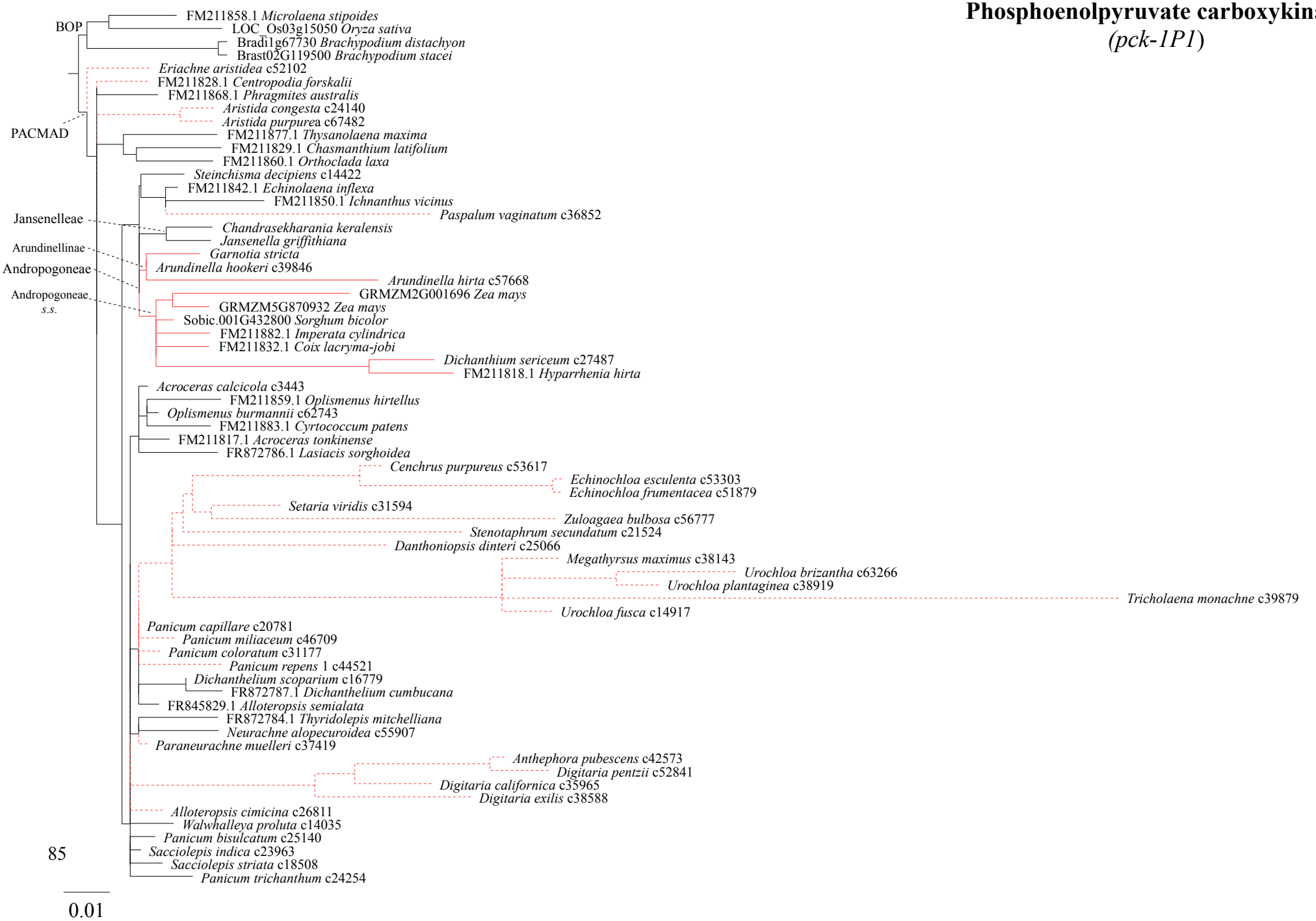
NADP-malate dehydrogenase
(nadpmdh-1P1)



Phosphoenolpyruvate carboxylase
(ppc-1P3)



Phosphoenolpyruvate carboxykinase
(pck-1P1)



Chapter 3.

Gene duplication and dosage effects during the early emergence of C_4 photosynthesis in the grass genus *Alloteropsis*

Chapter 3. Gene duplication and dosage effects during the early emergence of C_4 photosynthesis in the grass genus *Alloteropsis*

Matheus E. Bianconi, Luke T. Dunning, Jose J. Moreno-Villena, Colin P. Osborne and Pascal-Antoine Christin

Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK

This work was published in **Journal of Experimental Botany**, Volume 69, Issue 8, 9 April 2018, Pages 1967-1980.

Personal contribution: I performed the genomic analyses, generated genetic data and wrote the manuscript. All co-authors commented on the text before submission.

3.1. Abstract

The importance of gene duplication for evolutionary diversification has been mainly discussed in terms of genetic redundancy allowing neofunctionalization. In the case of C_4 photosynthesis, which evolved via the co-option of multiple enzymes to boost carbon fixation in tropical conditions, the importance of genetic redundancy has not been consistently supported by genomic studies. Here, we test for a different role for gene duplication in the early evolution of C_4 photosynthesis, via dosage effects creating rapid step changes in expression levels. Using genome-wide data for accessions of the grass genus *Alloteropsis* that recently diversified into different photosynthetic types, we estimate gene copy numbers and demonstrate that recurrent duplications in two important families of C_4 genes coincided with increases in transcript abundance along the phylogeny, in some cases via a pure dosage effect. While increased gene copy number during the initial emergence of C_4 photosynthesis probably offered a rapid route to enhanced expression, we also find losses of duplicates following the acquisition of genes encoding better-suited isoforms. The dosage effect of gene duplication might therefore act as a transient process during the evolution of a C_4 biochemistry, rendered obsolete by the fixation of regulatory mutations increasing expression levels.

Keywords: Biochemical pathway, C_4 photosynthesis, copy number variation, dosage effect, gene duplication, grasses, low-coverage sequencing

3.2. Introduction

C_4 photosynthesis is a complex trait that results from the co-ordinated action of multiple biochemical and anatomical components to concentrate CO_2 at the site of Rubisco, increasing photosynthetic efficiency under warm and dry conditions (Hatch 1987; Sage 2004). Despite its complexity, the C_4 trait evolved multiple times independently in several groups of angiosperms (Sage et al. 2011). All enzymes required for the C_4 pathway were present in non-C_4 ancestors, where they were responsible for different, non-photosynthetic functions (Sage 2004; Aubry et al. 2011). The evolution of C_4 photosynthesis consequently required the co-option of these enzymes into new functions, followed by changes in their expression patterns and/or catalytic properties (Bläsing et al. 2000; Tausta et al. 2002; Gowik et al. 2004; Akyildiz et al. 2007; Christin et al. 2007; Hibberd and Covshoff 2010; Huang et al. 2017). It has been hypothesized that this massive co-option was facilitated by gene duplication, with one of the duplicates acquiring the novel C_4 function via neofunctionalization while the other continued to fulfil the ancestral function (Monson 1999, 2003; Sage 2004). However, recent genomic studies have not supported this hypothesis of genetic redundancy facilitating neofunctionalization, meaning that the genomic mechanisms enabling the acquisition of novel functions during C_4 evolution remain largely unknown.

Most C_4-related enzymes are encoded by multigene families, with numerous paralogues that emerged via multiple rounds of whole-genome and single-gene duplications during angiosperm diversification (Wang et al. 2009; Christin et al. 2013, 2015; Huang et al. 2017). However, the number of paralogues within each of these gene families does not differ significantly between C_3 and C_4 species (Williams et al. 2012; van den Bergh et al. 2014). Comparative genomics on a handful of grasses have identified duplicates that have been retained on branches leading to two C_4 origins, but these did not encode enzymes necessarily involved in the C_4 cycle (Emms et al. 2016). Indeed, investigations focusing on gene families with a known function in C_4 photosynthesis indicate that the gain of a C_4-specific function was generally not directly preceded by a gene duplication event (Christin et al. 2007, 2009; Wang et al. 2009). Although the creation of a large reservoir of ancient duplications might still be important (Monson 2003), these various lines of evidence suggest that C_4 evolution did not consistently involve duplication followed by neofunctionalization of one copy while the other retained the ancestral function. However, gene duplication might still have

played a role in the initial emergence of C_4 photosynthesis, via a combination of dosage effects and neofunctionalization.

Small-scale or whole-genome duplications are generally expected to increase transcript abundance through a gene dosage effect (Otto et al. 1986; Kondrashov et al. 2002; Conant and Wolfe 2008; Conant et al. 2014). Instances of retention of duplicated genes due to a dosage effect on expression levels have been reported for a number of adaptive traits, which include insecticide resistance in the *Culex* mosquito (Mouchès et al. 1986), cold protection in Antarctic fishes (Chen et al. 2008), and nematode resistance in soybean (Cook et al. 2012). Positive selection on the dosage effect of newborn duplicates is predicted in cases where the protein products physically interact with molecules such as toxins or nutrients, or in cases in which proteins need rapid and constant production at high levels (Kondrashov et al. 2002; Kondrashov 2012). The dosage effect of gene duplication might consequently be important for the establishment of a C_4 cycle. Current models of C_4 evolution hypothesize that a weak C_4 cycle can first emerge using enzymes that have not been adapted to the C_4 catalytic context (Sage 2004; Heckmann et al. 2013; Christin and Osborne 2014; Mallmann et al. 2014; Heckmann 2016; Dunning et al. 2017). Gene duplications increasing the transcript abundance of C_4-related genes in plants with a weak C_4 cycle would increase the strength of the pathway, which is predicted to boost carbon assimilation and fitness (Heckmann et al. 2013; Mallmann et al. 2014), leading to the preferential retention of the duplicates. We propose here to test the hypothesis that gene duplications contributed to the initial emergence of a C_4 biochemistry via dosage effects, with subsequent neofunctionalization. We capitalize on the diversity of C_4 enzymes that evolved in the recent past within the grass genus *Alloteropsis*.

The *Alloteropsis* genus contains five species, four of which are C_4, while the fifth, *A. semialata*, encompasses C_4 as well as non-C_4 populations with and without a weak C_4 cycle (Ellis, 1974; Lundgren et al. 2016). The diversification of *A. semialata* took place during the last 3 million years (Lundgren et al. 2015), and only a few genes are markedly up-regulated in the C_4 accessions compared with C_3 populations (Dunning et al. 2017). In some cases, the identity of genes used for the C_4 cycle differs among C_4 populations of *A. semialata*, which is interpreted as the footprint of a gradual adaptation of C_4 photosynthesis during the diversification of the group involving secondary gene flow among previously isolated populations (Chapter 4; Dunning et al. 2017). This group therefore represents an outstanding system to investigate the small-scale

processes that led to C_4 photosynthesis, including the importance of genomic rearrangements such as duplications for C_4 evolution.

Genome scans coupled with genome size estimates are used here to assess the gene content of accessions of the genus *Alloteropsis* varying in their photosynthetic type, testing (i) whether the copy number of genes encoding C_4-related proteins varies among accessions of *Alloteropsis*; (ii) whether gene duplications coincide with the co-option of genes for a C_4 function; and (iii) whether increases in gene copy number result from the duplication of genomic material or from retroposition events (i.e. insertion of retrotranscribed RNA into the genome; Kaessmann et al. 2009). In addition, we retrieve published transcriptomes for members of the *Alloteropsis* genus (Dunning et al. 2017) and associate them with newly generated high-coverage genome sequencing to test (iv) whether recently duplicated genes are expressed; (v) whether multiple copies all contribute to overall transcript abundance; and (vi) whether increases in copy number of C_4-related genes along the phylogenetic tree were associated with increases in expression levels. This comparative analysis of gene copy numbers provides evidence for a potential role for recent gene duplications in physiological innovation through rapid and drastic changes of transcript abundance.

3.3. Materials and Methods

3.3.1. Taxon sampling and genome data

A total of 20 genome-wide, low-coverage sequencing datasets of *Alloteropsis* J. Presl were retrieved from published studies (Table 3.1; Lundgren et al. 2015; Chapter 4; NCBI accession no. SRP082653). These include two accessions of the C_4 *A. angusta* Stapf, one of the C_4 species *A. cimicina* (L.) Stapf, and 17 of *A. semialata* (R. Br.) Hitchc. Among these 17 *A. semialata*, 12 are C_4 individuals sampled across a broad geographical range from West Africa to Australia, and the five non-C_4 include three individuals with a weak C_4 cycle ('C_3+C_4' in Dunning et al. 2017; note that this term is equivalent to 'type II C_3–C_4 intermediates' *sensu* Edwards and Ku 1987) and two C_3 individuals from South Africa. Each of the genomic datasets consists of paired-end Illumina reads, with read lengths of 100, 125, or 150 bp (Table 3.1). In this study, the raw reads were filtered using the NGSQC Toolkit (Patel and Jain 2012) to retain only high-quality sequences (i.e. >70% of read length with Phred quality >20), and to remove primer and adaptor contaminated reads. The genome size and ploidy level of

some of the individuals analysed here were retrieved from previous studies that used the same accessions (Lundgren et al. 2015; Chapter 4). Some accessions were only available as herbarium samples, preventing estimates of genome sizes or ploidy levels.

High-coverage sequencing datasets were generated here for two individuals to allow single nucleotide polymorphism (SNP) analyses (see below). This included one C_3+C_4 accession from Tanzania (TAN2) already sequenced at low coverage and one C_4 accession from a population where another individual was sequenced at low coverage (TPE1; Table 3.1). For these two samples, 250 bp long paired reads were obtained with the Illumina technology.

The different sequence datasets were obtained from whole genomic DNA, so that reads can belong to any of the nuclear, chloroplast, and mitochondrial genomes. Reads from the two organellar genomes were identified by mapping the genomic datasets onto representative chloroplast and mitochondrial genomes using Bowtie2 (Langmead and Salzberg 2012) with default parameters, and removed before analyses. Mitochondrial genomes were assembled de novo (Supplementary text S1) using the approach described in Lundgren et al. (2015), while chloroplast genomes were retrieved from Lundgren et al. (2015) and Chapter 4. On average, 3% of the initial reads were removed because of their organellar origin (Table 3.1).

Table 3.1. Genome data information.

ID	Species	Carbon isotope	Genome size (Gb/2Cx[1])/Ploidy	Country	Transcriptome sample[2]	Sequencing batch[3]	Sequencer	Read Length	Total nuclear genome reads	Organellar reads (%)[4]	Theoretical Coverage[5]
Cim1	*A. cimicina*	C_4	-	Madagascar	ACIM	2	HiSeq 2500	100	20,898,025	2.0	0.95
Ang1	*A. angusta*	C_4	-	DRC	-	5	HiSeq 3000	150	14,751,007	2.6	1.01
Ang2	*A. angusta*	C_4	1.95 / 2n	Uganda	-	2	HiSeq 2500	100	18,665,954	1.9	0.96
RSA1	*A. semialata*	C_3	-	South Africa	-	1	HiSeq 2500	100	14,821,009	0.8	0.67
RSA2	*A. semialata*	C_3	1.80 / 2n	South Africa	KWT3	1	HiSeq 2500	100	12,524,356	0.6	0.70
TAN1	*A. semialata*	C_3+C_4	1.88 / 2n	Tanzania	LO4	2	HiSeq 2500	100	18,899,157	4.0	1.01
TAN2-A	*A. semialata*	C_3+C_4	2.19 / 2n	Tanzania	LO1	2	HiSeq 2500	100	20,065,838	4.2	0.92
TAN2-A[6]	*A. semialata*	C_3+C_4	2.19 / 2n	Tanzania	LO1	6	HiSeq 2500	250	45,774,384	3.4	5.05
TAN3	*A. semialata*	C_3+C_4	-	Tanzania	-	3	HiSeq 2500	125	35,782,290	1.6	2.03
DRC1	*A. semialata*	C_4	-	DRC	-	5	HiSeq 3000	150	33,933,832	3.6	2.31
DRC2	*A. semialata*	C_4	-	DRC	-	4	HiSeq 3000	150	23,098,686	3.1	1.57
DRC3	*A. semialata*	C_4	-	DRC	-	3	HiSeq 2500	125	28,889,427	6.4	1.64
DRC4	*A. semialata*	C_4	-	DRC	-	5	HiSeq 3000	150	14,749,392	4.0	1.01
TAN4	*A. semialata*	C_4	2.01 / 2n	Tanzania	LO2	2	HiSeq 2500	100	18,596,076	3.2	0.93
RSA3	*A. semialata*	C_4	5.22 / 6n	South Africa	MDB8	1	HiSeq 2500	100	13,824,190	0.8	0.26
KEN1	*A. semialata*	C_4	-	Kenya	-	3	HiSeq 2500	125	25,405,608	4.9	1.44
BUR1	*A. semialata*	C_4	1.95 / 2n	Burkina Faso	BF3	1	HiSeq 2500	100	13,498,418	0.9	0.69
MAD1	*A. semialata*	C_4	2.05 / 2n	Madagascar	MAJ	1	HiSeq 2500	100	16,440,692	1.8	0.80
THA1	*A. semialata*	C_4	-	Thailand	-	2	HiSeq 2500	100	16,873,534	2.1	0.77
TPE1-3	*A. semialata*	C_4	1.87 / 2n	Taiwan	TW10	2	HiSeq 2500	100	15,435,339	4.8	0.83
TPE1-10[6]	*A. semialata*	C_4	1.87 / 2n	Taiwan	TW10	7	HiSeq 2500	250	169,555,422	3.4	21.92
AUS1	*A. semialata*	C_4	2.20 / 2n	Australia	AUS2	1	HiSeq 2500	100	11,600,487	0.8	0.53

[1] Genome size (Gb/2Cx) = total genome mass (pg) x 0.978; [2] Data retrieved from Dunning et al. (2017; Table 3.S4); [3] Accessions with the same batch number were sequenced together; [4] Percentage of reads mapping to chloroplast and mitochondrial genomes; [5] Based on 2C genome size; after removing organellar reads; assuming a value of 2.2 Gb (maximum value of a diploid individual of *Alloteropsis*) for unknown genome sizes; [6] Dataset generated for this study. Other datasets were retrieved either from Lundgren et al. (2015) or Chapter 4.

3.3.2. Mapping of reads on reference datasets

Gene copy numbers were estimated using a modified read depth approach (Alkan et al. 2009; Yoon et al. 2009; Teo et al. 2012). This strategy divides the genome into non-overlapping regions (bins) and uses the number of genomic reads mapped to each of these regions to estimate gene copy number. Bins receiving in some accessions more or fewer reads than expected under a null statistical model are considered copy number variants (Fig. 3.1). Given the current lack of a reference genome for any *Alloteropsis* species, genomic data were mapped to a reference data set consisting of coding sequences (CDSs) of *A. cimicina* and *A. semialata*, which was retrieved from the transcriptome study of Dunning et al. (2017). Briefly, this data set comprises groups of co-orthologues at the Panicoideae subfamily level, the group of grasses that includes the genus *Alloteropsis*. Each group of co-orthologues encompasses all the genes that are descended by speciation and/or gene duplication from a single gene in the common ancestor of Panicoideae. Only genes captured in one of the *Alloteropsis* transcriptomes and with co-orthologues in at least one of *Sorghum bicolor* and *Setaria italica* were included. Increases in copy number detected here therefore correspond to duplications that happened after the initial diversification of Panicoideae, about 30 million years ago. Manually curated alignments using longer transcripts of 23 gene families with a known function in C_4 biochemistry (Bräutigam et al. 2011) and the gene encoding the Rubisco small subunit (*rbcS*) were added into the reference data set. These manually curated alignments improved read mapping accuracy in cases where paralogues with high sequence similarity were present, such as laterally acquired forms previously identified for phosphoenolpyruvate carboxylase (PEPC; *ppc* gene) and phosphoenolpyruvate carboxykinase (PCK; *pck* gene; Christin et al. 2012; Dunning et al. 2017). Overall, this genome-wide data set comprised 12688 groups of co-orthologues, belonging to 5589 gene families.

Genomic reads were mapped onto the genome-wide CDS data set using Bowtie2, with default parameters, randomly assigning reads mapped to multiple sequences to one of the top hits, and using the local alignment option. Reads were mapped as single-end reads to avoid false negatives when one of the reads mapped outside the CDS. The number of mapped reads (counts) per group of co-orthologues was obtained using SAMtools (Li et al. 2009) and used to compute gene copy number estimates as described below.

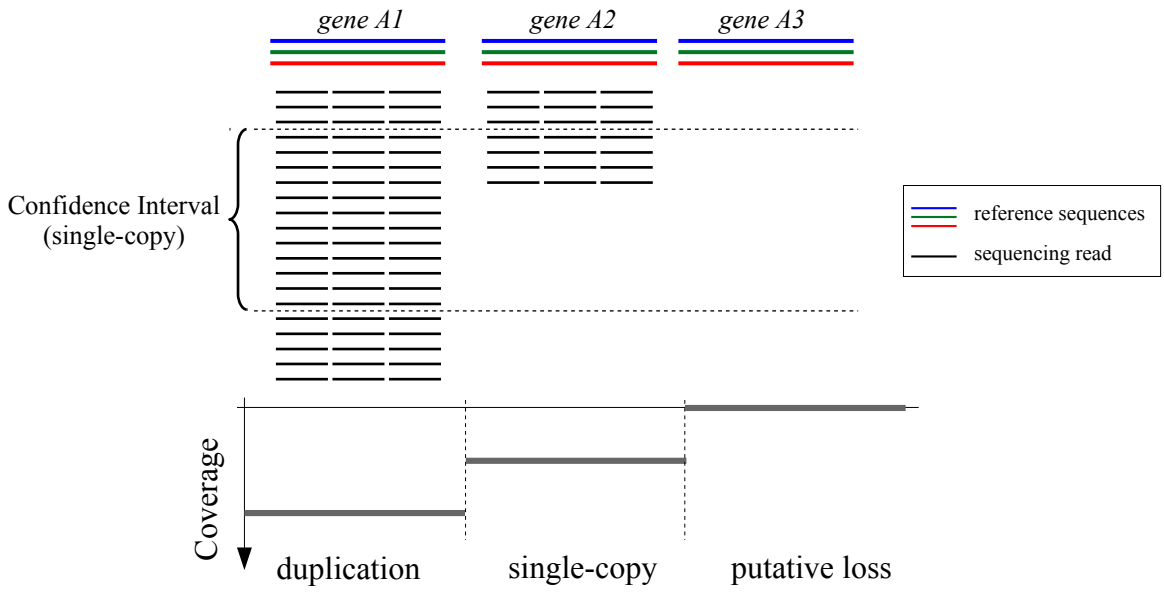


Fig. 3.1. Read depth approach for gene copy number estimation. Duplications are inferred when the number of read counts expected for a determined gene is significantly higher than the expected read counts for single-copy genes, according to an underlying statistical model.

3.3.3. Estimates of copy numbers

Under the assumption that each site in the genome has an equal probability of being the first site of a given read, the expected read count (c) for any genomic region i of length L can be computed as:

$$E(c_i) = N\,(L_i / G) \qquad (1)$$

where N is the total number of sequencing reads and G is the haploid genome size (in number of bases). Assuming the counts c is a random variable that follows a binomial distribution, with the total binomial trials being the total number of reads N, the probability of a region i being captured by one read is equivalent to the probability of success in each binomial trial, which is:

$$P = L_i / G \qquad (2)$$

A well-known complication of quantitative genomic studies based on read depth is the sequencing bias linked to the GC content of the sequenced region, which is particular to sequencing approaches where library preparation includes PCR steps, as

required for degraded DNA extracted from herbarium samples (Dohm et al. 2008; Aird et al. 2011; Benjamini and Speed 2012; Teo et al. 2012). The relationship between sequencing depth and GC content can vary across sequencing runs (Benjamini and Speed 2012), and previous studies have quantified this relationship using various metrics (Alkan et al. 2009; Bellos et al. 2012; Benjamini and Speed 2012). In this study, preliminary analyses confirmed that the relationship varied among the different batches of library preparation and sequencing (Fig. 3.S1). The relationship between read counts and GC content was consequently estimated for each sample by using the counts of genes extracted from the genome-wide reference mapping. Read counts were normalized by gene length, and genes with no count or counts > 1.5 times the median count were removed from this particular analysis, to enrich the data set with putative single-copy genes. These length-normalized counts were then expressed as a linear function of the mean GC content of the target genes (x_i), so that:

$$c_i / L_i = a + bx_i \quad (3)$$

The coefficients a and b were estimated individually for each genome data set using a linear model fit procedure in R (R Development Core Team 2017). To homogenize the number of genes across GC content classes, 60 genes were randomly drawn from those present in each of nine equally spaced classes of GC content from 38% to 78%, and linear coefficients were calculated on the pooled subsample. Only genes longer than 700 bp were used here, since such long genes receive more reads and therefore provide more accurate copy number estimates. This procedure was repeated 100 times, providing a non-parametric estimate of variation for the coefficients. An approximate correction of the binomial probability of success in each trial (Equation 2) by the GC content was then obtained by substituting Equations 3 and 1 in Equation 2, so that:

$$P = L_i \times (a + bx_i) / N \quad (4)$$

Note that these new probabilities are independent of the genome size and can therefore be estimated for any sample. If $E(c_i)$ is the expected count when a target gene is present as a single copy, an estimate of the absolute number of copies k_i can be obtained as:

$$k_i = c_i / \mathrm{E}(c_i) \quad (5)$$

The expected counts and confidence intervals for single-copy genes were computed using a binomial quantile function implemented in R, with a confidence level of 99% corrected for multiple comparisons using the Bonferroni method. Genes were considered duplicated if the counts were above the upper limit of this confidence interval, and single copy if the counts were within the confidence interval limits (inclusive). Although partial copies can exist following incomplete duplications, copy number estimates for duplicated genes were rounded up for follow-up analyses. Genes were considered absent when no read count was detected, provided the confidence intervals for the expected counts did not include zero. In such cases, and in cases where read counts were below the lower limit of the confidence interval, the genes were removed from the analysis, since accurate copy numbers could not be estimated.

3.3.4. Quantitative real-time PCR estimates of copy number

A number of concerns have been raised about the use of high-throughput sequencing data for genome analyses of structural diversity, such as copy number variants (Benjamini and Speed 2012; Teo et al. 2012). In particular, the above-mentioned GC content bias and others resulting from the library preparations represent potential caveats. We consequently performed quantitative real-time PCR (qPCR) assays to confirm the accuracy of the copy numbers estimated from the genome data. The gene family encoding the key C_4 enzyme phosphoenolpyruvate carboxylase (*ppc* genes) was selected for qPCR analyses since it included genes encompassing a wide range of copy numbers according to the read depth estimates (see the Results). Three paralogues (*ppc_1P3*, *ppc_1P6*, and *ppc_1P7*) were analysed in six individuals of *A. semialata* from a wide geographic and phylogenetic sampling (BUR1, RSA2, TAN2, TAN1, MAD1, and TPE1).

Alignments consisting of partial gene models of *ppc* groups of co-orthologues were assembled for *Alloteropsis* species using a genome-walking approach to include intron sequences, and were used as reference for primer design. Two pairs of primers per paralogue were designed to amplify 92–161 bp regions that include exon and intron sequences (except for one pair for *ppc_1P7*, which encompassed only exon sequences;

Table 3.S1). The copy number estimated via qPCR consequently captured only putative duplications of genomic DNA, and excluded potential retroposition instances (Zhang 2003; Kaessmann et al. 2009; Reams and Roth 2015). To perform the assays, genomic DNA (gDNA) was isolated from fresh leaves of *A. semialata* individuals using the DNeasy Plant Kit (Qiagen), following the manufacturer's instructions. SYBR green-based qPCRs were prepared using 1× Power SYBR green PCR Master Mix (Thermo Fisher Scientific), 0.25 µM of each primer, and 6.25 ng of gDNA in a total volume of 20 µl, with three technical replicates and non-template controls per reaction. Assays were carried out on a QuantStudio 12K Flex Real Time PCR instrument (Life Technologies) with an initial incubation of 10 min at 95 °C (Taq activation), followed by 40 cycles of 15 s at 95 °C (denaturation) and 60 s at 60 °C (annealing and extension). Amplification specificity was assessed via melting curves generated immediately after each assay, in which samples were incubated for 15 s at 95 °C and 60 s at 60 °C, followed by incremental temperature increases of 0.3 °C up to 95 °C. The melting temperature of the amplified fragments was then calculated based on their expected sequences and compared with the peak temperature values obtained from the melting curve assays. Baseline, threshold cycle, and PCR efficiency were determined using the LinRegPCR software v. 2016.0 (Ramakers et al. 2003). Samples with PCR efficiency <1.85 or >2.1 were excluded from the subsequent analysis. The Pfaffl method (Pfaffl 2001) was used to correct for different PCR efficiencies across amplicon groups, and copy numbers of *ppc* genes were expressed relative to the mean of the two pairs of primers used for the *ppc_1P7* gene.

3.3.5. Phylogenetic analyses of duplicated genes

To determine whether duplications of *ppc* and *pck* (see the Results) occurred before or after the diversification of *A. semialata* lineages, we assembled partial allele models by manually phasing polymorphisms using paired-end information. Ambiguous nucleotides were called for polymorphisms that could not be phased. Alleles of TPE1 and TAN2 were assembled using the high-coverage data, while raw transcriptome data of the genus *Alloteropsis* retrieved from Dunning et al. (2017) were used for the other accessions. Sequences were aligned using MAFFT v7.130b (Katoh and Standley 2013), and phylogenetic trees were inferred using PhyML (Guindon and Gascuel 2003) under a GTR+G model of nucleotide substitution, with 100 bootstrap pseudoreplicates.

3.3.6. Allele-specific expression analyses

The relative contribution of each allele/paralogue of *pck* and *ppc* to the overall transcript abundance was assessed and compared with their relative frequency in the genomes through the analysis of SNPs. Reads from the genome and transcriptome datasets were mapped to reference alignments of the *ppc* and *pck* gene families, and the read depth was determined for each SNP of each gene using Geneious v. 6.8 (Kearse et al. 2012). For each SNP, the abundance of the minor allele (defined on the transcriptome data as the variant base receiving fewer reads) was calculated as a proportion of the total read count for that site, for both transcriptome and genome data. Because the genomic frequency can vary among SNPs for multicopy genes (i.e. each variant can be present in any number of alleles up to twice the number of copies in a diploid individual), the contribution of different alleles to transcript abundance was evaluated via frequency correlations between transcriptome and genome datasets. Note that the polyploid individual was excluded from these analyses because of insufficient coverage to assess accurately polymorphisms among its high number of alleles.

3.3.7. Association between changes in copy number and transcript abundance

To test for an association between changes in copy number and changes in gene expression, transcript abundances in leaves were retrieved for 14 C_4-related genes captured in a study of transcriptomes of the genus *Alloteropsis* grown in controlled conditions (Dunning et al. 2017). The average abundance between two biological replicates in reads per kilobase per million mapped reads (RPKM) is used here. Values were log10 transformed before analysis to homogenize variances. Accessions were considered for this analysis only if genome and transcriptome data were available for the same individual, or individuals from the same population, except in two cases (representing *A. cimicina* and the C_3 *A. semialata*) for which genome and transcriptome data were available for closely related individuals from different populations (Lundgren et al. 2015; Chapter 4). Note that excluding these two individuals did not significantly alter the results. High-coverage sequence data were not used here to avoid pseudoreplication of some populations.

Homologous genes within a gene family do not represent independent data points as they result from events of gene duplication and/or speciation from a common

ancestor. We consequently used phylogenetic generalized least squares (PGLS) under a Brownian model of evolution to test for correlated changes between gene copy number and transcript abundance using the R packages nlme and APE (Paradis et al. 2004). A Bonferroni correction was used to adjust significance levels for multiple testing. The sequence alignment of the respective gene family was extracted from the genome-wide data set generated from transcriptomes (see above), and the accessions with no associated genome data were removed. Bayesian trees were inferred from this alignment under a GTR+G+I substitution model using MrBayes v3.2.2 (Ronquist et al. 2012), with two parallel analyses running for 10000000 generations. After verifying the convergence of the runs, a consensus tree was generated using trees sampled after a burn-in period of 50%. The effect of topological uncertainty on the PGLS results was assessed by repeating the analysis using 100 independent trees sampled every 50000 generations after the burn-in period.

3.4. Results

3.4.1. Background distribution of gene copy numbers

Copy numbers were estimated for markers sampled across the genome for each accession, providing a background distribution of copy numbers per haploid chromosome set (Fig. 3.S2). Most genes were estimated as single copy, and the proportion of duplicated genes ranged from 9% to 28% across accessions, with 0.5–1.3% genes being absent (Table 3.2). The same copy numbers were estimated among individuals belonging to the same nuclear group, as previously defined in *A. semialata* (Chapter 4), for 82% of the genes, on average. Although there was a weak positive correlation between coverage and the proportion of absent genes (R^2=0.34, P=0.055), no significant association was found between coverage and the proportion of single-copy (R^2=0, P=0.41) or duplicated genes (R^2=0, P=0.53), which suggests that the inferred duplications reflect biological rather than methodological differences. Similar estimates were found moreover between individuals from the same population based on low- and high-coverage datasets (Fig. 3.S3), indicating that low-coverage sequencing provides an accurate assessment of gene copy number variation. The variation in genome size (Table 3.1) was not explained by differences in gene copy number, with correlations being non-significant for the proportion of both absent and duplicated genes.

Table 3.2. Background distribution of gene copy numbers in *Alloteropsis* accessions.

Accession	Species	Photosynthetic metabolism	Total genes analysed[1]	Proportions (%)[2]		
				Single-copy	Duplicated	Absent
Cim1	*A. cimicina*	C_4	12,057	89.4 (88.2 - 90.6)	9.8 (8.6 - 11)	0.9 (0.8 - 0.9)
Ang1	*A. angusta*	C_4	8,966	83.9 (81.4 - 86.9)	14.8 (11.7 - 17.4)	1.2 (1.2 - 1.4)
Ang2	*A. angusta*	C_4	9,700	84.2 (81.6-85.8)	14.5 (12.8 - 17.1)	1.3 (1.3 - 1.4)
RSA1	*A. semialata*	C_3	8,935	83.8 (81.7 - 85.9)	15.5 (13.4 - 17.6)	0.7 (0.7 - 0.8)
RSA2	*A. semialata*	C_3	6,996	86.4 (84.9 - 88)	13.1 (11.3 - 14.6)	0.5 (0.5 - 0.7)
TAN1	*A. semialata*	C_3+C_4	11,376	88.4 (87.5 - 89.3)	10.8 (9.9 - 11.6)	0.8 (0.8 - 0.9)
TAN2	*A. semialata*	C_3+C_4	11,221	86.1 (85.3 - 87.2)	13.2 (12.1 - 14.1)	0.7 (0.7 - 0.7)
TAN3	*A. semialata*	C_3+C_4	12,195	79.5 (77.7 - 82.2)	19.9 (17.2 - 21.7)	0.6 (0.6 - 0.6)
DRC1	*A. semialata*	C_4	12,162	79 (76.4 - 81.3)	20.4 (18.1 - 23)	0.6 (0.6 - 0.6)
DRC2	*A. semialata*	C_4	11,946	81.1 (78.5 - 83.1)	18.3 (16.3 - 20.9)	0.6 (0.6 - 0.6)
DRC3	*A. semialata*	C_4	11,941	78.3 (75.3 - 80.7)	21 (18.6 - 24)	0.7 (0.7 - 0.7)
DRC4	*A. semialata*	C_4	11,014	81.4 (79.1 - 83.9)	17.9 (15.4 - 20.2)	0.7 (0.6 - 0.7)
TAN4	*A. semialata*	C_4	11,214	86.6 (85.6 - 87.3)	12.6 (11.8 - 13.6)	0.8 (0.8 - 0.8)
RSA3	*A. semialata*	C_4	10,248	88.1 (86.3 - 89.4)	11.2 (9.9 - 13.1)	0.6 (0.6 - 0.7)
KEN1	*A. semialata*	C_4	10,381	70.6 (64.1 - 76.5)	28.4 (22.5 - 35)	1 (1 - 1)
BUR1	*A. semialata*	C_4	9,448	88.4 (87.4 - 89.5)	10.9 (9.7 - 11.9)	0.7 (0.7 - 0.8)
MAD1	*A. semialata*	C_4	10,226	88.1 (86.7 - 89.1)	11.2 (10.2 - 12.6)	0.7 (0.7 - 0.7)
THA1	*A. semialata*	C_4	10,926	87.5 (86 - 88.6)	11.7 (10.6 - 13.3)	0.8 (0.7 - 0.8)
TPE1	*A. semialata*	C_4	10,730	88.5 (87.5 - 89.3)	10.7 (9.9 - 11.7)	0.8 (0.7 - 0.8)
AUS1	*A. semialata*	C_4	7,174	88.3 (87 - 89.7)	11 (9.6 - 12.3)	0.7 (0.6 - 0.7)

[1] After removing genes with confidence intervals for the expected read counts including zero, and genes with read counts between 1 and and the lower limit of the confidence interval (See methods); [2] Percentage of single-copy, duplicated or absent genes relative to the total number of genes analysed. Values are medians calculated from the resampling procedure used for the GC-content correction, with the minimum and maximum values shown between parentheses.

3.4.2. Duplications of C_4 protein-coding genes

We estimated copy numbers for a total of 82 genes belonging to 23 gene families with some gene lineages encoding proteins known to be involved in the C_4 pathway of some species. For 45 of these genes belonging to 19 families, at least one duplication was observed in the genus *Alloteropsis* (Table 3.S2). Putative ancient duplications (shared by *A. semialata*, *A. angusta*, and *A. cimicina*) include those for pyruvate kinase (*pk_1P1*) and NADP-dependent malic enzyme (*nadpme_1P4*). A number of genes have incurred independent duplications and/or secondary losses within *A. semialata* and *A. angusta*, including those for a tonoplast malate/fumarate transporter (*tdt_1P2*), in addition to those encoding phosphoenolpyruvate carboxylase (*ppc_1P3*) and phosphoenolpyruvate carboxykinase (*pck_1P1_LGT:C*). The *pck_1P1_LGT:C* gene was laterally acquired after the split between the C_3 lineage and the lineage including C_3+C_4 and C_4 *A. semialata*, which now use it as part of their C_4 cycle (Chapter 4; Dunning et al. 2017), and subsequently duplicated only in the C_4 group (Fig. 3.2). The *ppc* gene family has a particularly high diversity of copy numbers, which is especially marked for *ppc_1P3* and *ppc_1P6*, both of which are used for the C_4 cycle of some accessions (Dunning et al. 2017).

The phylogenetic distribution of duplicates could be explained by different combinations of duplications and secondary gene losses (Fig. 3.2), but these scenarios can be distinguished based on gene trees. The multiple copies of *pck_1P1_LGT:C* retrieved from the C_4 *A. semialata* form a monophyletic clade, which is split into subgroups corresponding to African and Asian/Australian accessions (Fig. 3.S4). This pattern could be explained by independent duplications in each of the two groups or a duplication at their base followed by recombination or concerted evolution within each of the groups. The multiple copies of *ppc_1P6* specific to TPE1 (and THA1; Fig. 3.2), which is the only accession to use this gene for its C_4 pathway (Dunning et al. 2017), are very similar and cluster in the phylogeny (Fig. 3.S5), which supports the hypothesis of very recent duplications. The multiple *ppc_1P3* copies of the C_3+C_4 and C_4 *A. semialata* form distinct, well-supported monophyletic groups and, within the C_4 group, copies from the same accession tend to cluster despite a lack of resolution in some parts of the tree (Fig. 3.S6). This, again, suggests either independent duplications or concerted evolution following early duplications. Secondary losses of extra copies of *ppc_1P3* and the complete loss of *ppc_1P6* are inferred in the Australian accession (AUS1), which is

the only accession carrying one of the laterally acquired *ppc* genes (*ppc_1P3_LGT:A*; Fig. 3.2).

The copy numbers estimated for *ppc_1P3* and *ppc_1P6* from the genome data were significantly correlated with those estimated by qPCR (R^2=0.88, P<0.001; Fig. 3.3). Since intronic regions were amplified in both pairs of primers used for the qPCR analysis, we conclude that the observed duplications correspond to duplications of genomic DNA. Differences in copy number of *ppc_1P3* between different primer pairs may be explained by the existence of a polymorphism in a region amplified by one of the primers, which would prevent the amplification of one of the alleles. Analyses of sequence alignments confirmed this was the case for at least one individual (MAD1). Alternatively, it is also possible that in other accessions some of the duplicates are present as partial copies originating from illegitimate recombination.

3.4.3. Increases in transcript abundance associated with lineage-specific duplications

Our analyses of C_4-related genes revealed remarkable variation in copy number of *ppc* and *pck* among *Alloteropsis* lineages. For each polymorphic site, the frequency of the minor variant was strongly correlated between high-coverage genome and transcriptome datasets across the eight copies of *ppc_1P6* identified in TPE1 by the qPCR analysis (R^2=0.93, P<0.001; Fig. 3.4). While the correlation between transcriptome and genome sequencing was also observed for *ppc_1P3* of TPE1, it was weaker (R^2=0.38, P=0.06; Fig. 3.4; Table 3.S4), which might stem from lower overall transcript abundance and a small number of SNPs increasing statistical noise, or variation in the transcript contribution of different copies. The association between genome and transcriptome SNP frequencies varied among the other samples (Table 3.S4), which reflects a combination of low genome coverage of individual variants, variants not shared among the individuals used for genome and transcriptome sequencing, and biased transcriptome contribution of different copies. Nonetheless, the analyses of *ppc_1P3* and *pck_1P1_LGT:C* genes clearly show that multiple copies are expressed at consequent levels in the C_3+C_4 and C_4 accessions, contributing to the elevated overall transcript levels of these genes in the C_3+C_4 and C_4 *A. semialata* (Table 3.S3; Dunning et al. 2017). Overall, the SNP analyses provide strong support for duplicates being equally expressed in some accessions (e.g. *ppc_1P6* of TPE1), and show a widespread

contribution of multiple copies to the elevated transcript abundance of *ppc* and *pck* genes.

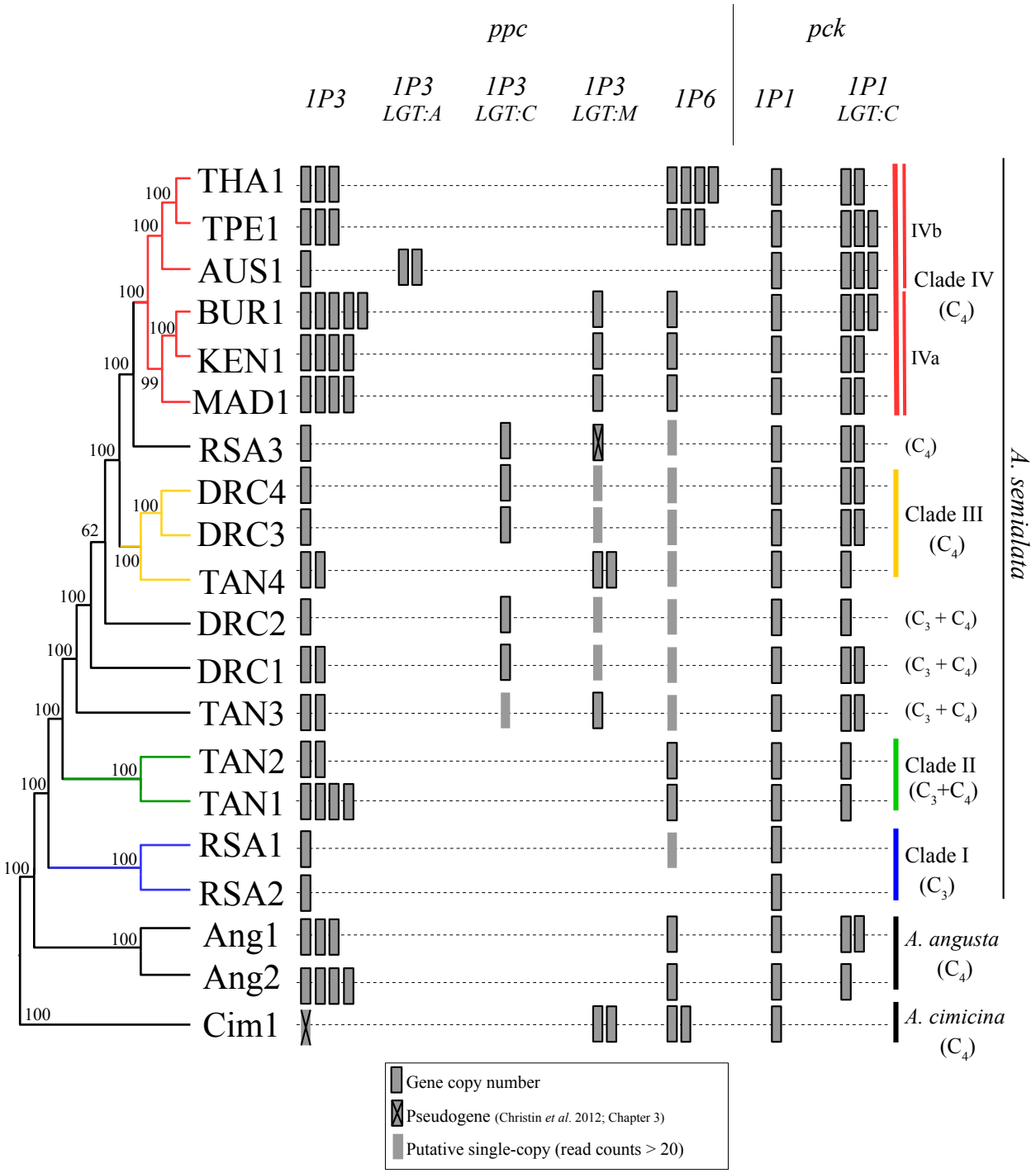


Fig. 3.2. Copy number variation of selected genes of phosphoenolpyruvate carboxylase (*ppc*) and phosphoenolpyruvate carboxykinase (*pck*) in the *Alloteropsis* genus. LGT:A, C, and M are laterally acquired genes (Christin et al. 2012). Nuclear phylogeny of the *Alloteropsis* genus was modified from Chapter 4, with lineages indicated. Copy number estimates are based on low-coverage genome data.

Finally, we tested whether the observed changes in copy number were statistically associated with changes in transcript abundance during the evolutionary diversification of the genus *Alloteropsis*. The conclusions of the statistical tests are robust to topological uncertainty (Table 3.S5), and we therefore discuss here only the results of the PGLS analyses based on the consensus tree (Table 3.3). Out of the 14 C_4-related gene families for which transcript abundance was available in Dunning et al. (2017), 10 showed copy number variation among the accessions used for this analysis. We found a consistent positive association between changes in copy number and changes in transcript abundance that was significant after correction for multiple testing in two of them, *ppc* (P<0.001) and *pck* (P=0.002; Table 3.3; Fig. 3.5; Fig. 3.S7). In the case of *ppc*, these effects were mainly driven by a few copy number changes in *ppc_1P3* and *ppc_1P6* (Fig. 3.5A), which, along with the laterally acquired *ppc* genes (*ppc_1P3_LGT:A*, *ppc_1P3_LGT:M*, and *ppc_1P3_LGT:C*), are the most highly expressed copies of this gene family in the C_4 accessions of the *Alloteropsis* genus (Dunning et al. 2017). For *pck*, the duplication of *pck_1P1_LGT:C* after the split between the C_3+C_4 and C_4 lineages was tightly associated with increases in transcript abundance of this gene (Fig. 3.5B). Although the other eight families include, in some cases, genes varying in copy number and transcript abundance, the statistical association was not significant after taking the phylogeny into account. In addition, analyses of *rbcS* showed a decrease in abundance in C_3+C_4 and C_4 accessions, which was associated with increases in gene copy numbers, highlighting processes other than dosage effects during the diversification of this gene family in terms of copy number and transcript abundance (Fig. 3.S8).

Table 3.3. Test for the association between changes in gene copy number and changes in transcript abundance in *Alloteropsis*.

Gene family	Copy number range	Transcript abundance range[1]	*p*-value[2]
Alanine aminotransferase (ALA-AT)	1 – 2	0 – 1,838	0.08
Aspartate aminotransferase (ASP-AT)	1 – 2	9 – 2,632	0.48
Carbonic anhydrase (CA)	1 – 3	3 – 13,169	0.46
Dicarboxylate transporter (DIT)	1	0 – 342	-
NAD-malate dehydrogenase (NAD-MDH)	1 – 4	21 – 1,528	0.11
NAD-malic enzyme (NAD-ME)	1 – 2	12 – 162	0.57
NADP-malate dehydrogenase (NADP-MDH)	1	15 – 3,537	-
NADP-malic enzyme (NADP-ME)	1 – 3	0 – 5,746	0.56
PEP carboxykinase (PCK)	1 – 3	11 – 5,187	**0.002**
PEP carboxylase (PEPC)	1 – 5	0 – 11,153	**< 0.001**
Pyruvate phosphate dikinase (PPDK)	1 – 2	0 – 12,796	0.82
PEP-phosphate translocator (PPT)	1 – 2	19 – 2,593	0.62
Sodium bile acid symporter (SBAS)	1	17 – 7,105	-
Triosephosphate-phosphate translocator (TPT)	1 – 2	8 – 3,213	-

[1] In RPKM; retrieved from Dunning et al. (2017); [2] *p*-values were obtained using a phylogenetic generalized least squares (PGLS) fitting under a Brownian model of character evolution. Gene families lacking *p*-values do not show copy number variation, or contain representatives with no gene sequence available for the phylogenetic analysis.

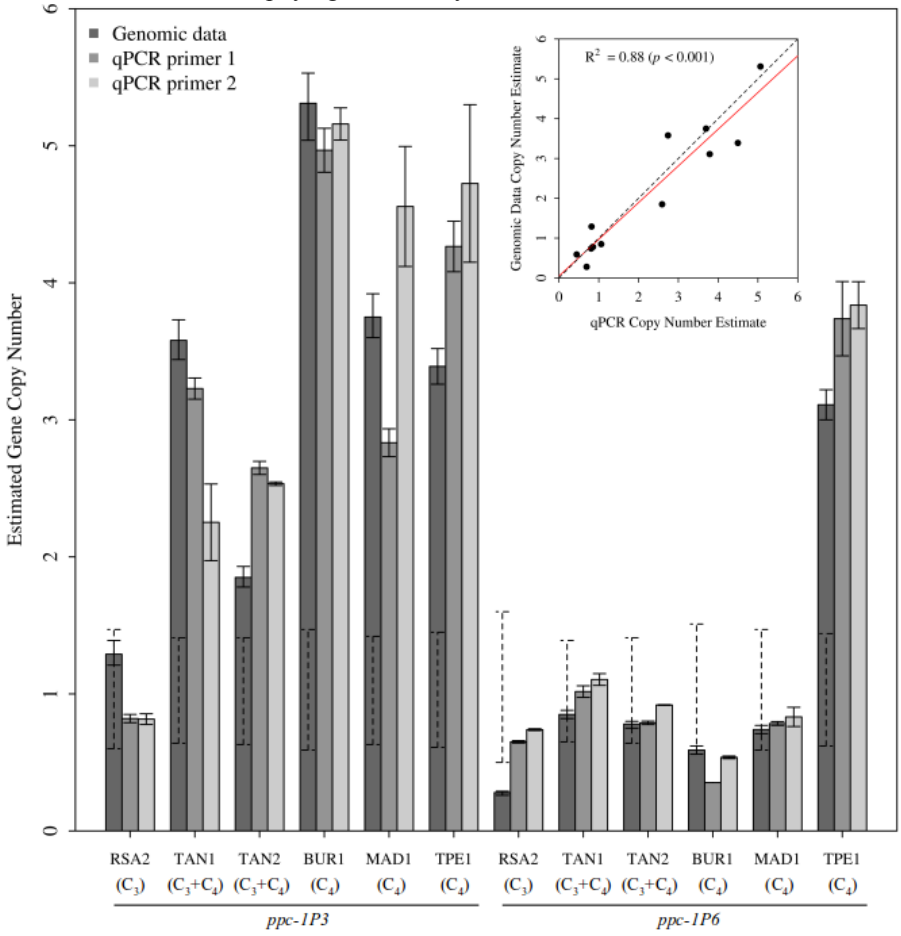


Fig. 3.3. Comparison of copy number estimates obtained from qPCR assays and from low-coverage genomic data for the genes *ppc_1P3* and *ppc_1P6* in six *A. semialata* accessions. Copy numbers are expressed relative to the *ppc_1P7* gene. Error bars are SEs from 2–3 technical replicates for qPCR estimates, and non-parametric error estimates from the GC correction resampling procedure for the genomic estimates of copy number. Dashed error bars for the genomic estimates are confidence intervals for expected single-copy genes. The upper panel indicates the correlation between qPCR estimates (mean value of both pairs of primers) and genomic estimates of copy number for *ppc_1P3* and *ppc_1P6*, with the solid line being the regression line and the dashed line the identity line.

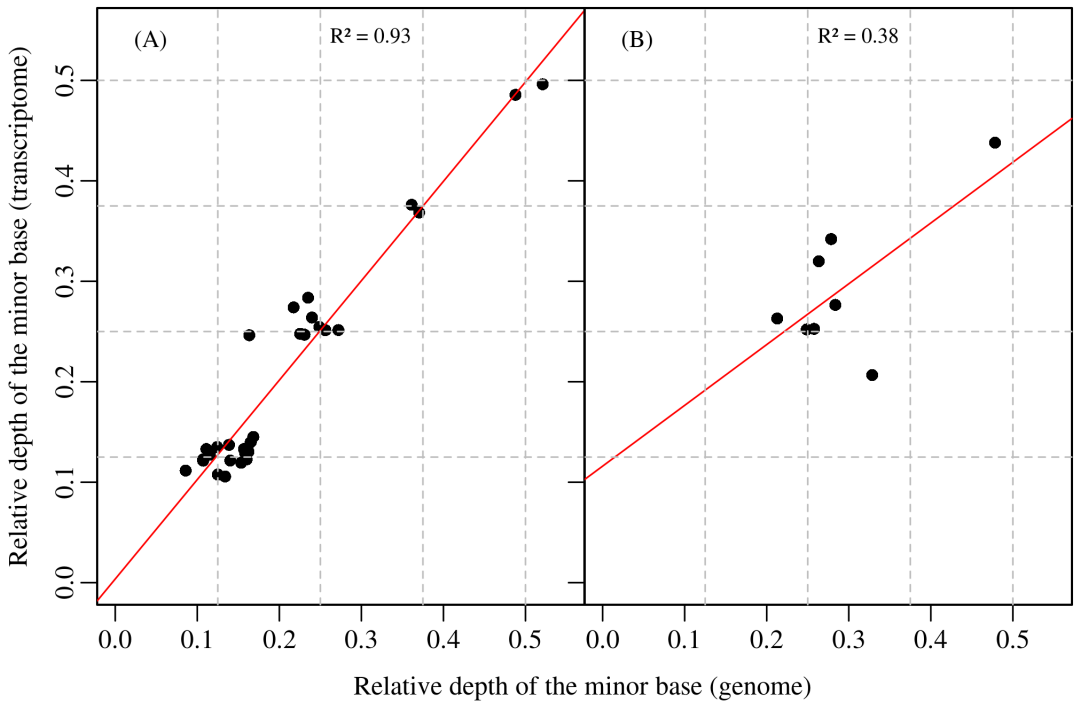


Fig. 3.4. Relative read depth of variants detected at polymorphic sites of (A) *ppc 1P6* and (B) *ppc 1P3* genes in the genome and transcriptome of a C_4 individual of *A. semialata* (TPE1). Each data point is a polymorphic site and is expressed as the depth of the minor base relative to the total depth for that site. The red line is the fitted linear model of transcriptome and genome data and the dashed black line is the identity line. The data points cluster around frequencies of 0.125, 0.25, 0.375, and 0.5, which correspond to one, two, three, and four alleles out of a total of eight alleles from four duplicates.

3.5. Discussion

3.5.1. Recent gene duplications linked to physiological innovation via potential dosage effects

In this study, we used genome analyses to show that genes for *ppc* and *pck* recurrently increased in numbers during the evolution of C_4 photosynthesis in the genus *Alloteropsis* (Fig. 3.2). These genes encode some of the few enzymes that reach very high levels in the C_3+C_4 and C_4 *A. semialata* (Ueno and Sentoku 2006; Lundgren et al. 2016; Dunning et al. 2017), and increases in copy numbers statistically coincided with enhanced transcript abundance (Table 3.3; Fig. 3.5). One potential explanation for this pattern is that increased gene expression and high transcript abundance favoured frequent retroposition; that is, high transcription caused gene duplication (Kaessmann et al. 2009). However, if this were the case, we would expect that increased copy number would uniquely involve exon sequences, which is disproved by our qPCR results. Analyses of polymorphisms further demonstrate that the multiple copies contribute to the overall high transcript abundances, with at least in some cases an equal contribution

from each copy (Fig. 3.4). We therefore conclude that duplication of genomic DNA directly contributed to the expression levels of these genes, via dosage effects. Modifications of the regulatory mechanisms during the diversification of land plants and grasses are probably responsible for the variation of transcript abundance observed among single-copy gene lineages, and recent duplications would then have quickly enhanced the transcript level associated with some of the ancestral gene (Fig. 3.5), which can reach consequent levels in the non-C_4 ancestors (Moreno-Villena et al. 2018). Evidence for this mechanism was obtained here for only two genes, which encode proteins that are responsible for the initial fixation of atmospheric carbon into organic compounds and the release of CO_2 to feed the C_4 cycle, respectively. Three other enzymes show marked increases in transcript abundance in the C_3+C_4 and/or C_4 *A. semialata* (Dunning et al. 2017), without evidence of gene copy number increases (Table 3.3). Unsurprisingly, the proposed dosage effect therefore concerns only a subset of the C_4 genes, but it probably played a key role first in the emergence of a weak C_4 cycle in the C_3+C_4 accessions, and then in the strengthening of this cycle in the C_4 accessions, which is predicted to impact positively on fitness (Heckmann et al. 2013; Mallmann et al. 2014; Bräutigam and Gowik 2016). Our results therefore suggest that dosage effects contributed to physiological innovation in the studied taxa, in association with changes in the regulatory properties of genes encoding other enzymes.

Establishing the context of the duplications behind these increased copy numbers would require assembled genomes, but could involve unequal crossing over, chromosomal duplication, or the action of transposable elements (Zhang 2003; Reams and Roth 2015). Using high-coverage sequencing from genomic DNA or transcriptomes, we were able to assemble multiple copies of some *ppc* and *pck* genes in diploid accessions of *A. semialata*. While phylogenetic trees supported early duplications in some cases, the copies tended to group per accessions (Figs 3.S4–3.S6). The number of assembled copies was moreover below that estimated based on sequencing depth, suggesting that identical alleles exist. These patterns could be explained by recurrent gene duplications during the history of the *Alloteropsis* genus, or recombination, for example among tandem duplicates, leading to concerted evolution homogenizing the duplicated copies within geographically isolated lineages (Brown et al. 1972; Nei and Rooney 2005).

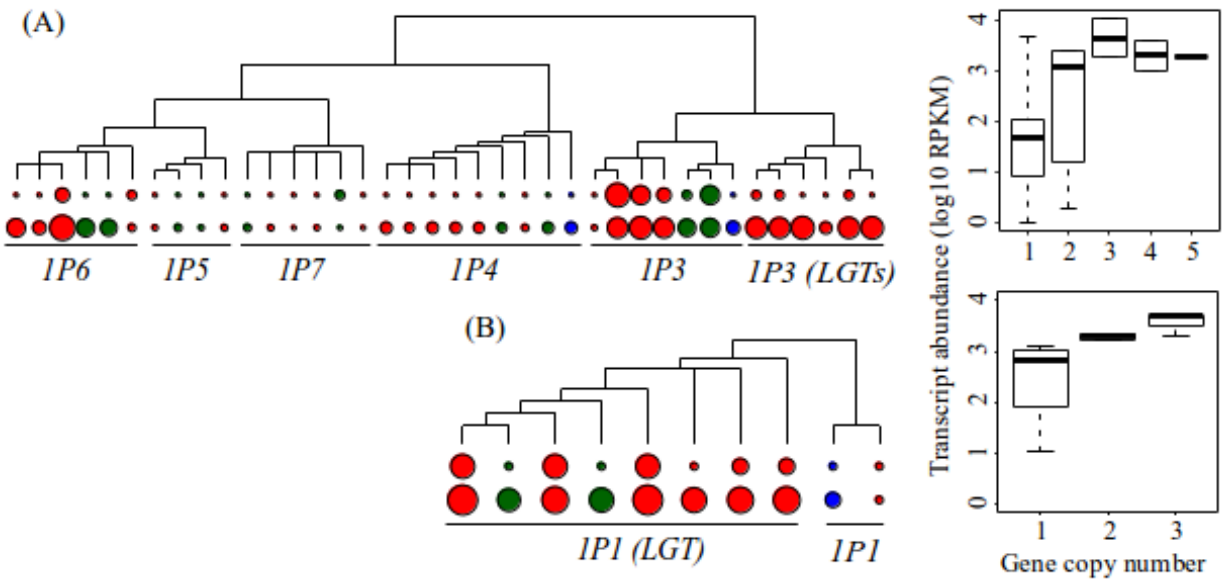


Fig. 3.5. Association between changes in gene copy number and transcript abundance for (A) phosphoenolpyruvate carboxylase (*ppc*) and (B) phosphoenolpyruvate carboxykinase (*pck*). For each gene in each accession, circles next to the tips of the gene phylogeny are proportional to the estimated gene copy number (top) and transcript abundance (log10 RPKM; bottom). Dots are coloured according to the photosynthetic type (blue=C_3, green=C_3+C_4, red=C_4). The boxplots on the right show the distribution of transcript abundances per class of copy numbers.

3.5.2. Duplicates get lost after the acquisition of better-suited copies

At least three events of lateral gene transfers (LGTs) of *ppc* and one of *pck* occurred in the *Alloteropsis* genus (Christin et al. 2012; Chapter 4), and some of the laterally acquired genes are expressed at high levels in the transcriptome of the accessions carrying such genes (Dunning et al. 2017). In most of these accessions, the vertically inherited copies of *ppc* and *pck* are strongly down-regulated, or not expressed at all (Dunning et al. 2017). Apart from the Southeast Asian clade, all C_4 accessions of *A. semialata* studied here carry at least one laterally acquired *ppc* gene in their genomes. Interestingly, in this exception, multiple duplications of *ppc_1P6* were retained and are associated with drastic changes in transcript abundance that are specific to this clade (Fig. 3.5). On the other hand, the presence of some LGT copies (*ppc_1P3_LGT:C* and *ppc_1P3_LGT:A*) coincides with the loss of the initial duplicates of the vertically inherited *ppc_1P3* gene (Fig. 3.2; Chapter 4). These findings indicate that, once a gene better suited for the C_4 function is acquired, the selective pressure on the original copy is relaxed, leading over time to pseudogenization and/or gene loss.

With multiple copies of genes related to C_4 metabolism, the chances that some of these copies will acquire C_4 adaptive mutations increase. Our analyses indeed identified non-synonymous polymorphisms among multiple copies of some genes. In four cases, such substitutions on *ppc* generate amino acid changes that were recurrently selected in a number of other C_4 grasses, suggesting that they adapt the protein for the C_4 catalytic context (Christin et al. 2007). While not detectable with our approach, regulatory mutations, identified for other C_4 groups (e.g. Gowik et al. 2004; Akyildiz et al. 2007), might similarly be present in only some of the multiple copies reported here. Genes that do not have the adaptive mutations can be lost via negative selection or drift, and those with the beneficial mutations are retained, leading to typical neofunctionalization. As reported here, the acquisition of more suitable gene versions, illustrated by the LGTs, can indeed relax the selection over duplicated copies that were once preserved via dosage selection, but from there on will be subjected to pseudogenization or eventually neofunctionalization. This suggests that during the course of evolution, fewer, more optimized genes are likely to remain, which would explain why more established C_4 lineages are not enriched in C_4-related genes (Williams et al. 2012; van den Bergh et al. 2014). The presence of multiple gene copies therefore probably contributes to the emergence of C_4 photosynthesis via a combination of dosage effects and increased opportunities for neofunctionalization, both of which are evolutionarily transient.

3.5.3. Low-coverage sequencing correctly identified duplicates

Low-coverage genomic datasets are increasingly used for a wide range of population genomic (Buerkle and Gompert 2013; Nicod et al. 2016; Chapter 4) and phylogenetic studies (Bock et al. 2014; Dodsworth 2015; Washburn et al. 2015). While such datasets are relatively cheap to obtain and can be generated from poorly conserved samples such as those from museum collections (Besnard et al. 2014; Silva et al. 2017), they come with their limitations. In particular, sequencing biases are inherent to the PCR steps involved in the sample preparation, and lead to over-representation of regions with specific GC contents (Benjamini and Speed, 2012; Ross et al. 2013; see the Materials and methods). It is therefore necessary to validate the results with independent evidence, provided here by qPCR. Slight variation between qPCR estimates and those based on low-coverage data confirmed that copy numbers inferred from read depths are in some cases under- or overestimated, as expected given both the low coverage and the

difficulty in precisely correcting for the sequencing bias. However, the general patterns are correctly identified, as indicated by the similarity of estimates among closely related accessions, and by the strong agreement in the estimates based on low- and high-coverage datasets in cases where both were available for individuals from the same population (Fig. 3.S3). In addition, individual events of gene duplication inferred from low-coverage data are qualitatively correct, being in all cases confirmed by independent qPCR.

The intersection of different lines of evidence shows that our approach represents a valid strategy to infer patterns of copy number variation for a large number of non-model species. Some of the genomic datasets included here come from samples only available in herbarium collections, which were collected up to 60 years ago (Chapter 4). In cases where living material is not available, low-coverage sequencing represents a valuable resource to shed light not only on the phylogenetic relationships, but also the genomic content of important taxa (Besnard et al. 2014), and, as shown here, variation in gene copy number. In the near future, the increasing availability of sequencing datasets for non-model species will offer multiple opportunities to track the genomic dynamics underlying a large array of physiological adaptations in a variety of taxa.

3.6. Conclusions

Using comparative genomics, we showed that the duplication of genes encoding two key enzymes required for C_4 photosynthesis coincided with the co-option of these genes for the new metabolic pathway. Based on published transcriptome data, we propose that changes in copy number altered the expression levels via pure dosage effects, with duplication events representing major effect mutations that can rapidly double transcription levels of some genes, which might have contributed to the emergence of a weak C_4 cycle in some plants. Once the C_4 cycle was in place, selection could act to optimize it, which probably involved fixing beneficial mutations on individual genes, including substitutions and indels in both regulatory and coding sequences. The selection of better-suited isoforms apparently led to pseudogenization of the previous duplicates. We therefore suggest that gene copy number decreases as beneficial mutations in the promoter or coding sequences are fixed, in a process of neofunctionalization. The beneficial effects of gene duplication for physiological

innovation are therefore likely to be transitory, with no footprint on longer evolutionary scales.

3.7. Acknowledgements

We thank Ilia Leitch and Oriane Hidalgo for comments on genome size results. MEB is supported by the Brazilian Research Council (CNPq) through a 'Science without Borders' scholarship (grant no. 201873/2014-1), LTD by an NERC grant (grant no. NE/M00208X/1), JJMV by a Royal Society Research Grant (grant no. RG130448), and PAC by a Royal Society University Research Fellowship (grant no. URF120119). The laboratory work was supported by the UK Natural Environment Research Council (NERC) Biomolecular Analysis Facility at the University of Sheffield. Library preparation and sequencing were carried out by Edinburgh Genomics, The University of Edinburgh. Edinburgh Genomics is partly supported through core grants from the NERC (R8/H10/56), MRC (MR/K001744/1), and BBSRC (BB/J004243/1).

3.8. Supporting Information

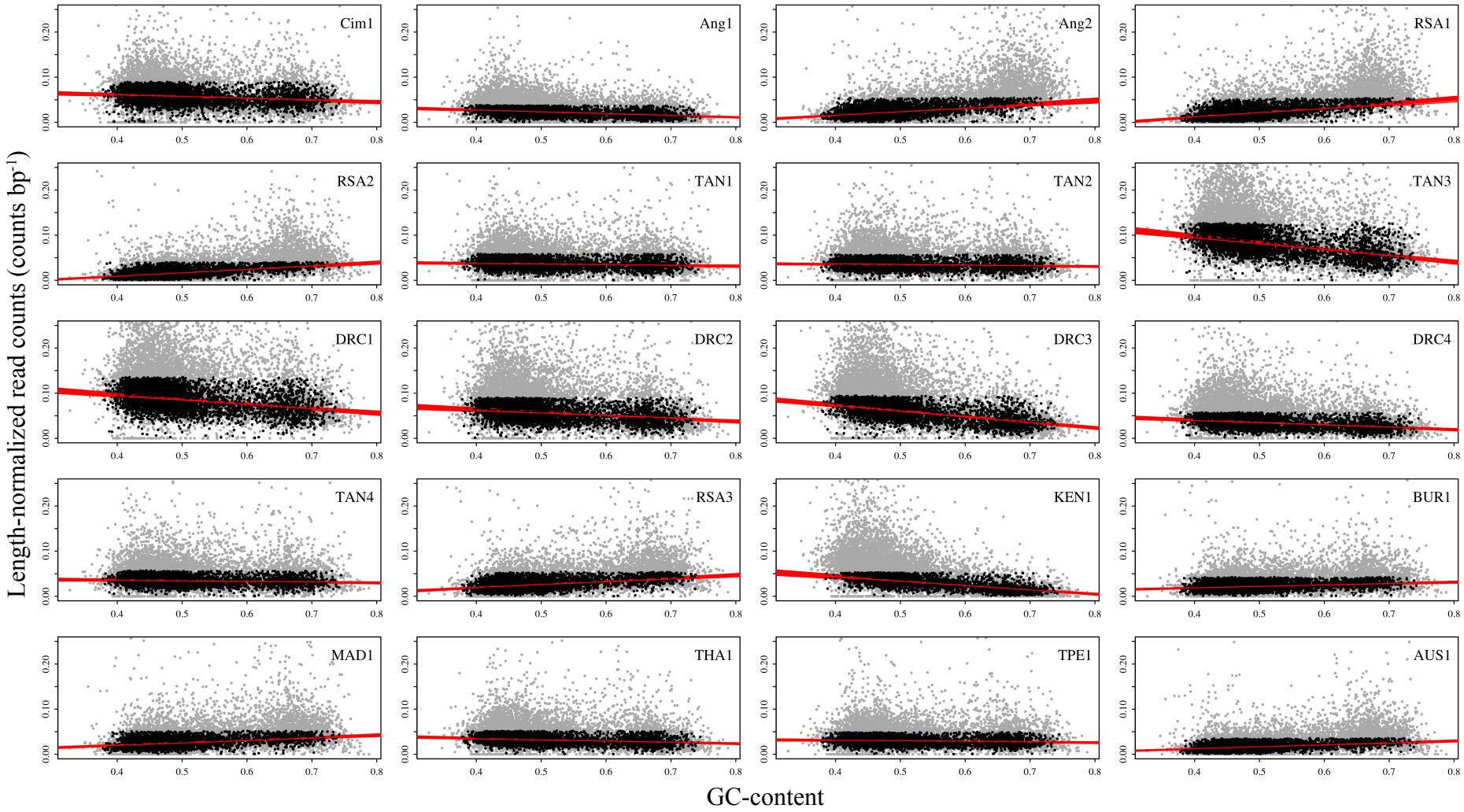


Fig. 3.S1. Relationship between length-normalized read count and GC-content in the genomic datasets of accessions of the genus *Alloteropsis*. Black points fall in the interval presumably enriched in single-copy genes that were randomly resampled for the nonparametric error estimation of gene copy numbers, and gray points are the whole set of genes analysed in this study. Red lines represent the linear regression of length-normalized read count and GC-content for each set of resampled genes.

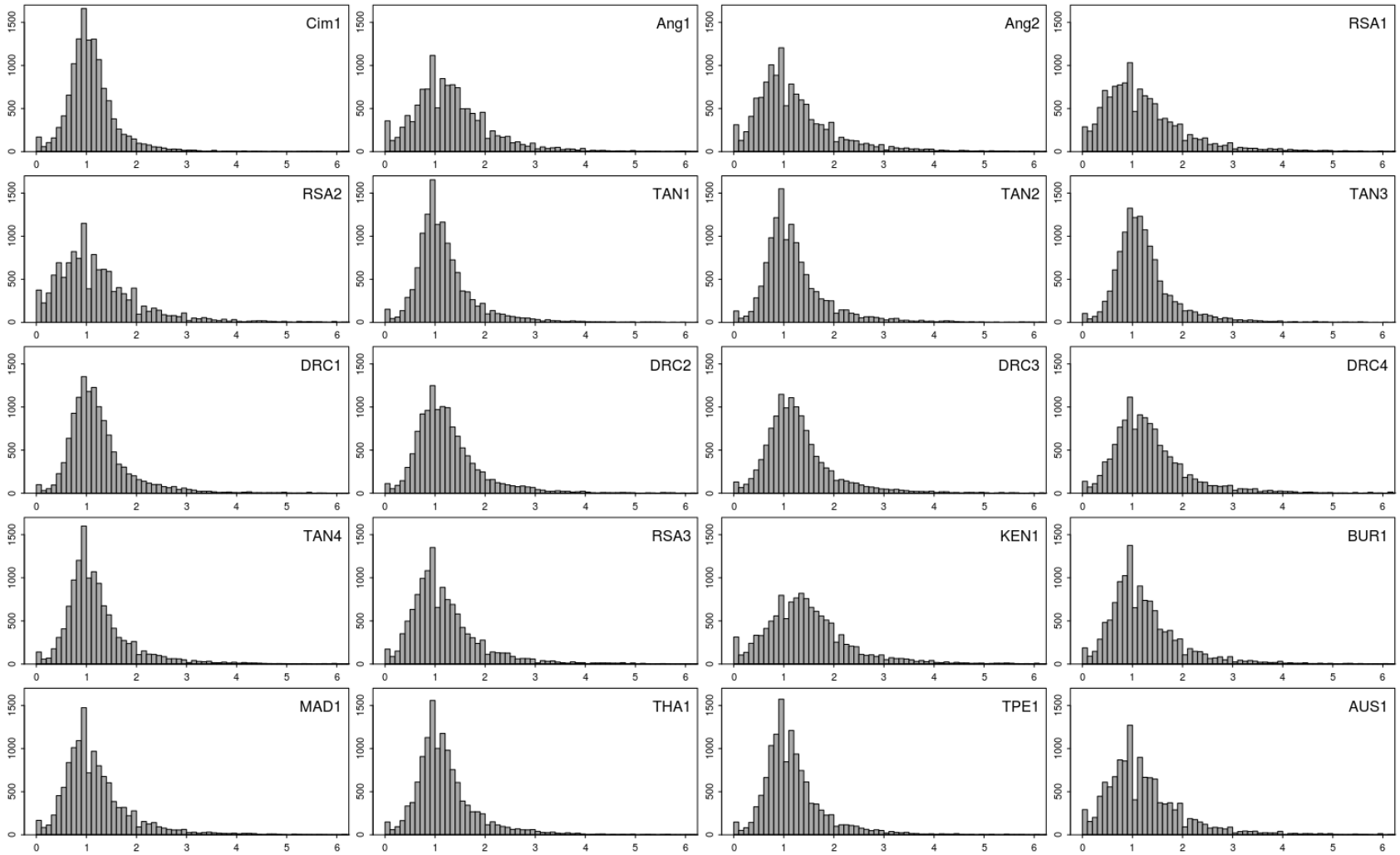


Fig. 3.S2. Background gene copy number distribution in accessions of the genus *Alloteropsis*. Copy numbers are expressed as observed read count divided by expected read count.

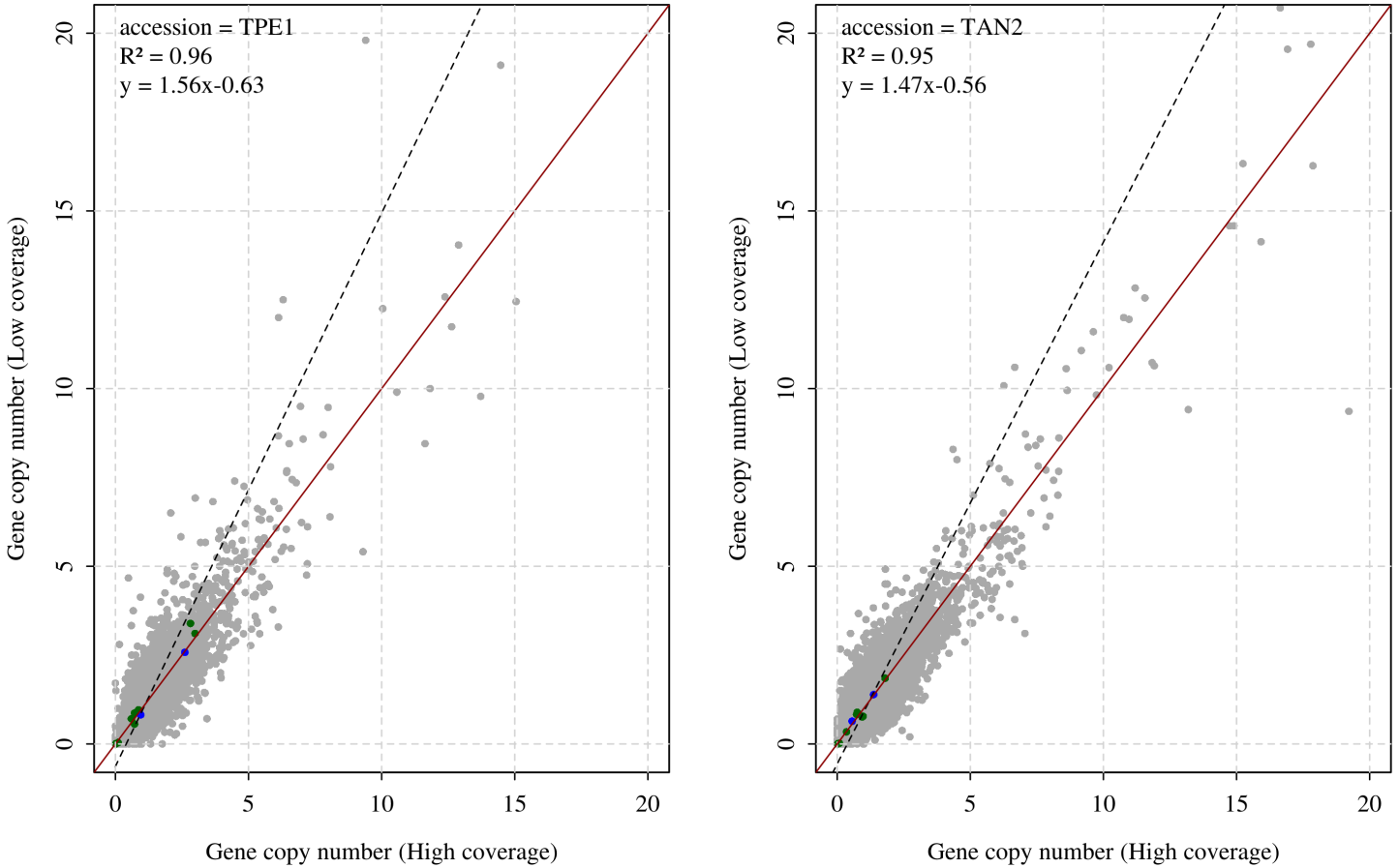


Fig. 3.S3. Comparison between copy number estimates using high coverage and low coverage datasets for individuals within the same population. The black dashed lines represent the linear regressions of low coverage (y) and high coverage estimates (x), and the red lines indicate identity. Coloured points are copy number estimates for the gene families *ppc* (green) and *pck* (blue).

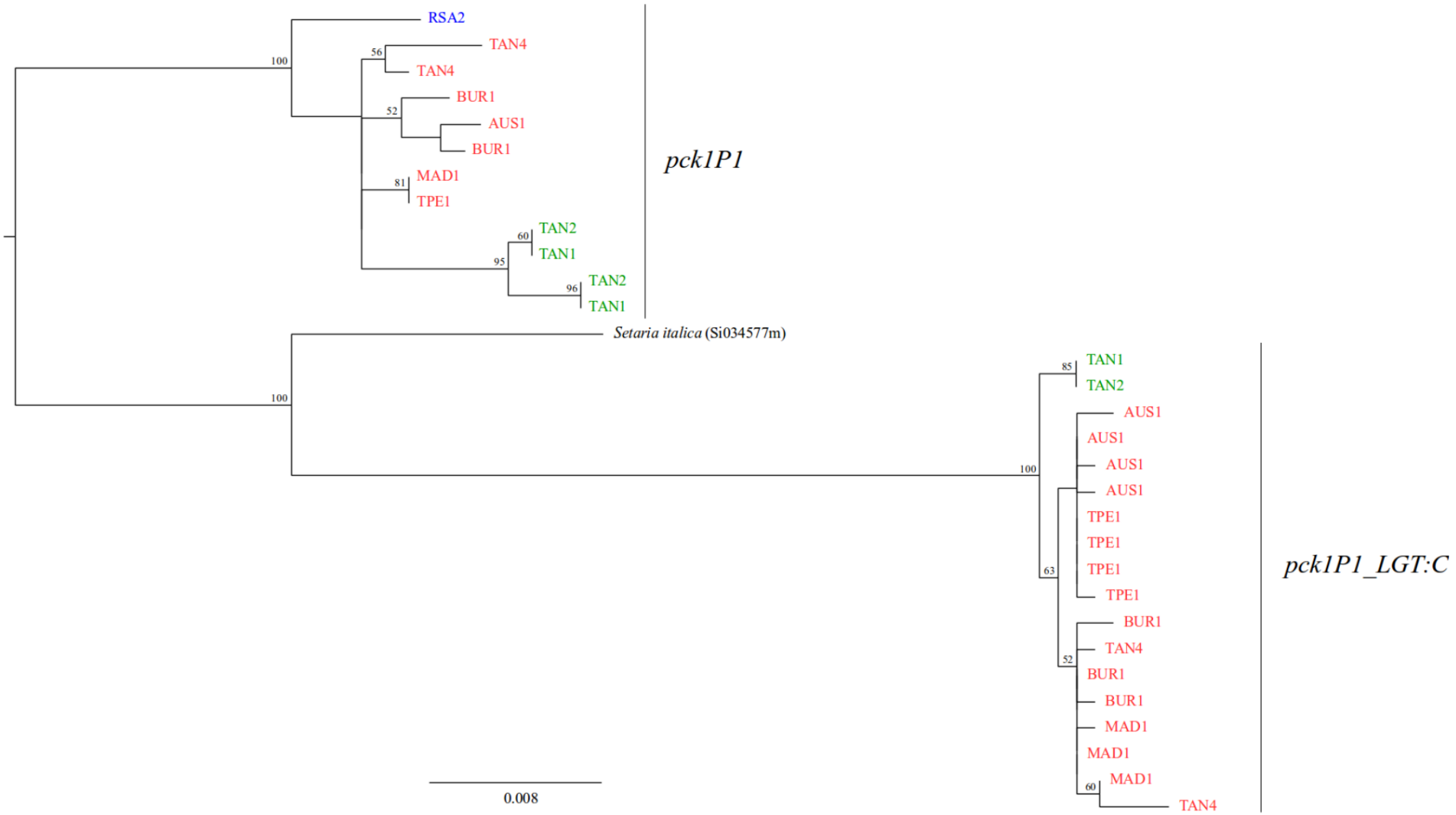


Fig. 3.S4. Phylogenetic tree of *pck* genes in the genus *Alloteropsis*. Colours indicate C_3 (blue), C_3+C_4 (green) and C_4 (red) accessions of *A. semialata*. Bootstrap support values are shown near branches when greater than 50.

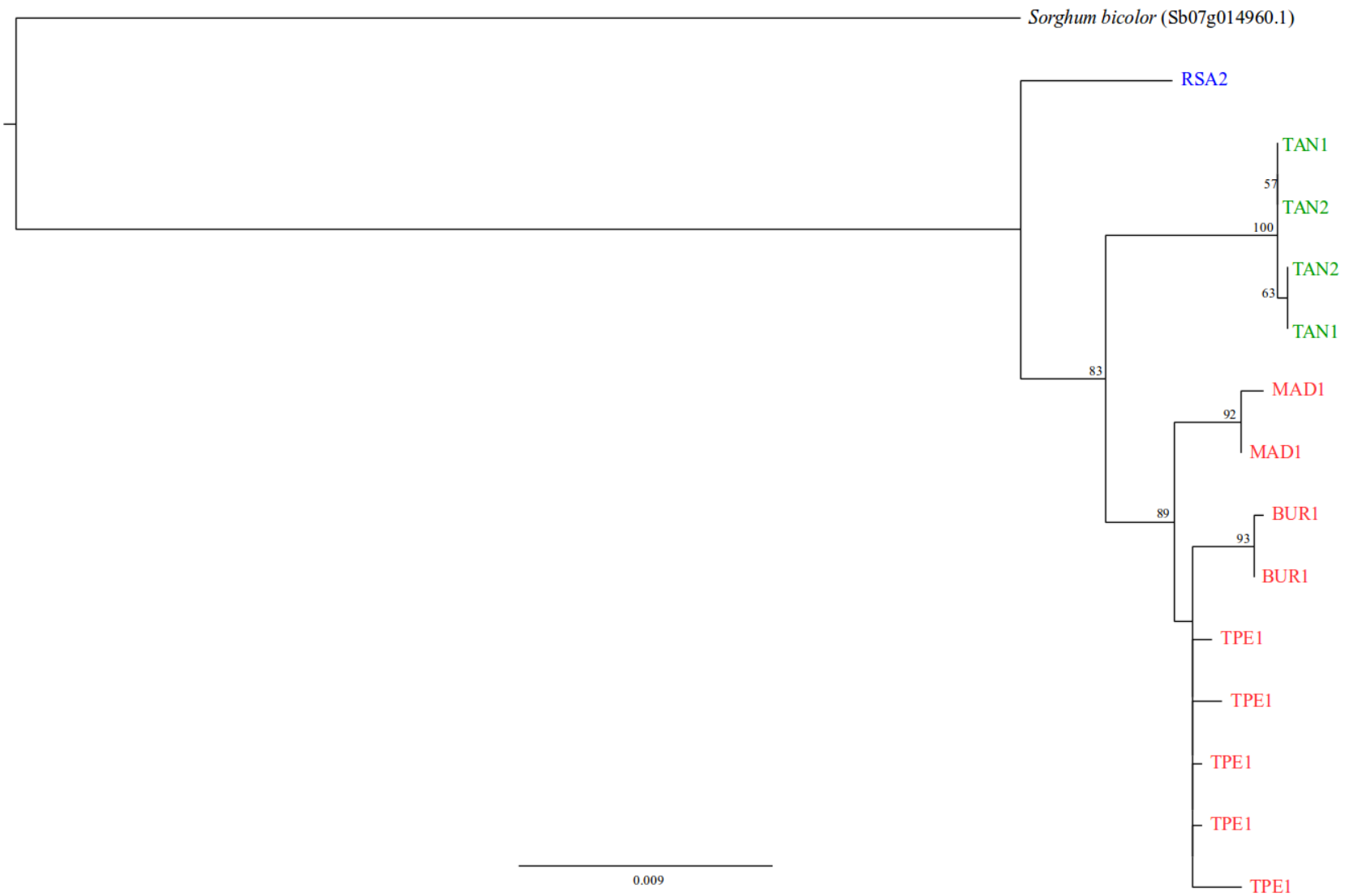


Fig. 3.S5. Phylogenetic tree of *ppc 1P6* genes in the genus *Alloteropsis*. Colours indicate C_3 (blue), C_3+C_4 (green) and C_4 (red) accessions of *A. semialata*. Bootstrap support values are shown near branches when greater than 50.

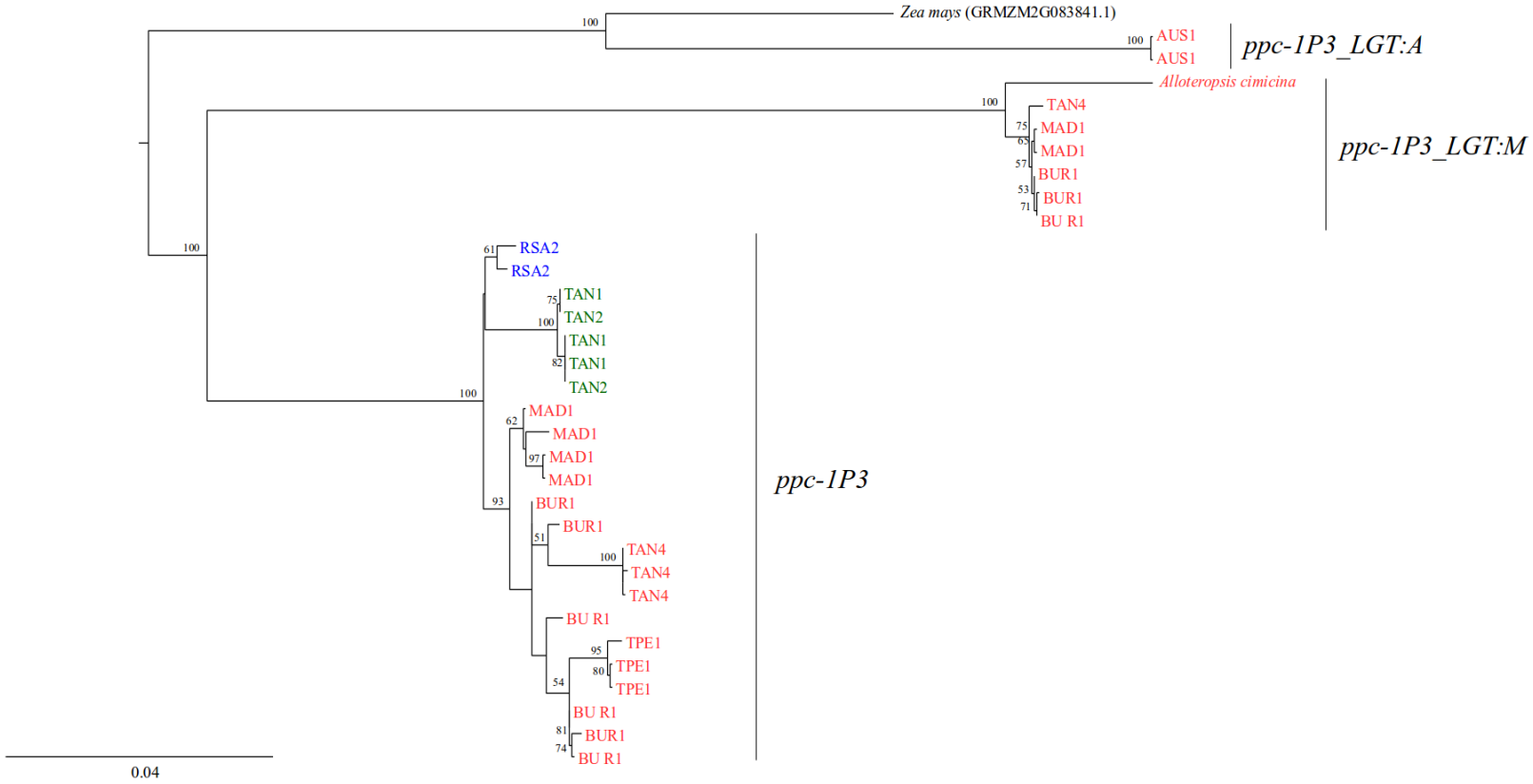


Fig. 3.S6. Phylogenetic tree of *ppc_1P3* genes in the genus *Alloteropsis*. Colours indicate C_3 (blue), C_3+C_4 (green) and C_4 (red) accessions of *A. semialata*. Bootstrap support values are shown near branches when greater than 50.

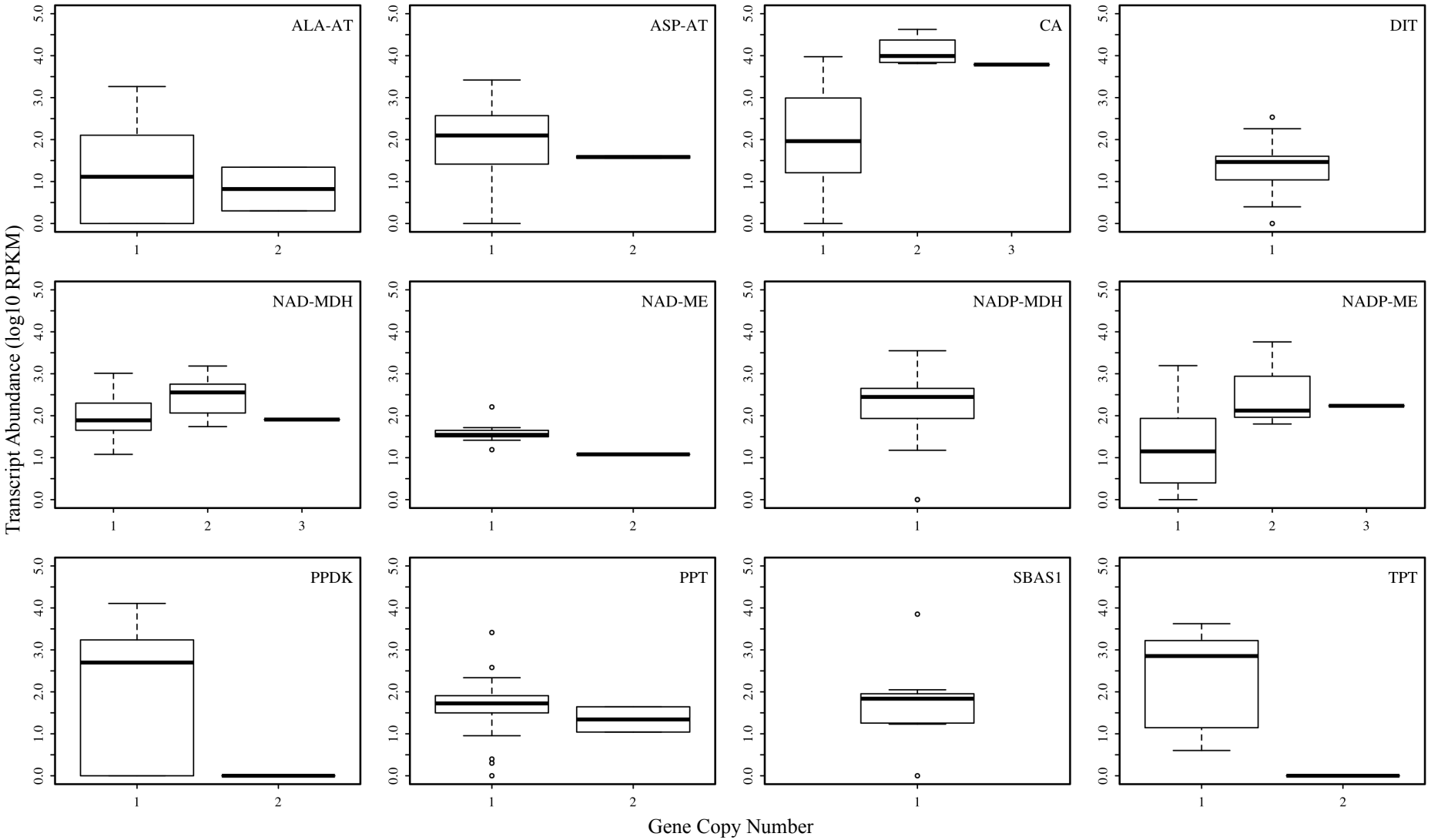


Fig. 3.S7. Distribution of transcript abundance among classes of gene copy numbers for 12 C_4-related gene families.

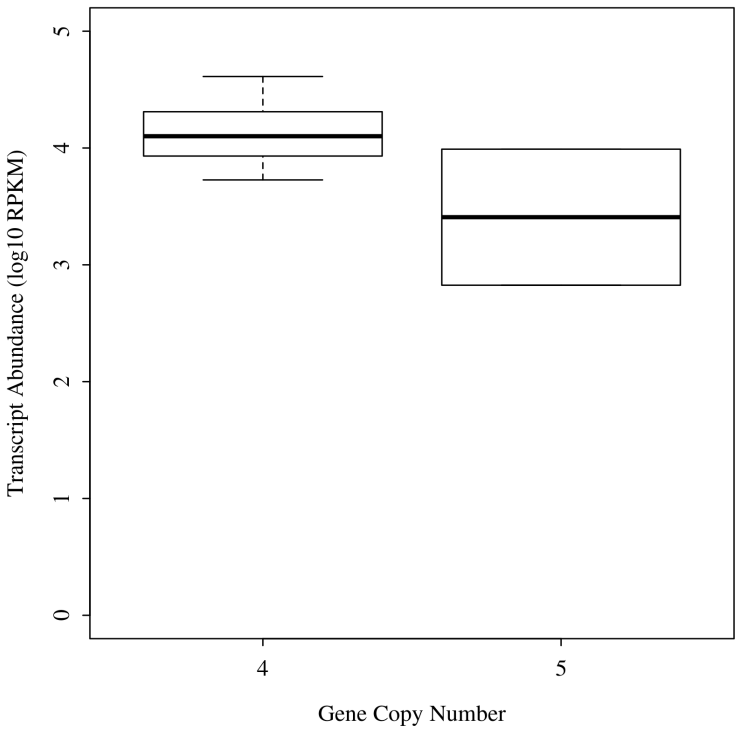


Fig. 3.S8. Distribution of transcript abundance among classes of copy numbers for genes encoding the small unit of ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco; *rbcS*).

Table 3.S1. List of primer sequences of *ppc* genes used for quantitative real-time PCR assays.

Gene	Primer ID	Sequence	Expected amplicon length/ Melting temperature
ppc_1P6 - pair 1 (exon 6 – intron 6)	ppc1P6-FOR-p1	5'-GCACGAGCAGATGAATTAC-3'	92 bp 73.8°C
	ppc1P6-REV-p1	5'-GTGAAAGTAGCCATCTACCATG-3'	
ppc_1P6 - pair 2 (exon 7 – intron 7)	ppc1P6-FOR-p2	5'-GCACGCCAGTTGTTATCCAG-3'	109 bp 76.4°C
	ppc1P6-REV-p2	5'-CAATATGTGCAGTTCAAAGGTTC-3'	
ppc_1P3 - pair 1 (exon 8 – intron 8)	ppc1P3_native-FOR-p1	5'-GTTCGTCGAGTACTTCCGATC-3'	115 bp 78.8°C
	ppc1P3_native-REV-p1	5'-GTGTGGCCTGACACGATC-3'	
ppc_1P3 - pair 2 (exon 8 – exon 9)	ppc1P3_native-FOR-p2	5'-CGGTTCGTCGAGTACTTCCG-3'	136 bp 80.4°C
	ppc1P3_native-REV-p2	5'-CGTACTCCGTCTCAGGTGTGG-3'	
ppc_1P7 - pair 1 (exon 7 – intron 7)	ppc1P7-FOR-p1	5'-CGTGTGATTCTGAGTGATGTC-3'	161 bp 78.3°C
	ppc1P7-REV-p1	5'-GCTAGACAAATCGAATGACCAC-3'	
ppc_1P7 - pair 2 (exon 8)	ppc1P7-FOR-p2	5'-CCGACATACTGATGTTATGGA-3'	119 bp 76.7°C
	ppc1P7-FOR-p2	5'-ACGGCCTCTTTCCATTAAGTT-3'	

Table 3.S2. List of duplicated genes of C_4-related gene families within *Alloteropsis*.

Gene family	Gene	*Alloteropsis* lineages with gene duplications[1]
Adenylate kinase (AK)	*ak_1P1*	I
	ak_2P2	I, IVa
Alanine aminotransferase (ALA-AT)	*alaat_1P2*	II, III, IVa
	alaat_1P3	II, III, IVa, MB
	alaat_1P5	III, IVa, MB
Aspartate aminotransferase (ASP-AT)	*aspat_1P1*	II, III, MB
	aspat_1P2	IVa
	aspat_2P3	Ang, MB
	aspat_3P4	Ang, III, IVa, MB
Carbonic anhydrase (CA)	*ca_1P1*	MB
	ca_2P2	III
	ca_2P3	Cim, II, III, IVa, IVb, MB
Dicarboxylate carrier (DIC)	*dic_1P1*	Cim, II, III, IVa, MB
Dicarboxylate transporter (DIT)	-	no duplications
Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)	*gapdh_2P1*	II, III, IVa, IVb, MB
	gapdh_2P2	IVa
	gapdh_3P1	Ang
NAD-dependent malate dehydrogenase (NAD-MDH)	*nadmdh_1P1*	II, IVb
	nadmdh_1P8	Cim
	nadmdh_2P4	III
	nadmdh_3P5	II, III, IVa, IVb, MB
NAD-malic enzyme (NAD-ME)	*nadme_1P1*	Cim, IVa
	nadme_2P2	IVa
NADP-dependent malate dehydrogenase (NADP-MDH)	*nadpmdh_1P1*	IVa, MB
NADP-malic enzyme (NADP-ME)	*nadpme_1P1*	Cim, III, IVa, IVb, MB
	nadpme_1P2	Ang, IVa
	nadpme_1P4	Cim, Ang, III, IVa, MB
PEP carboxykinase (PCK)	*pck_1P1_LGT:C*	Ang, III, IVa, IVb, MB
PEP carboxylase kinase (PEPC-K)	*pepck_1P3*	Ang
	pepck_3P6	IVa
Pyruvate kinase (PK)	*pk_1P1*	Cim, Ang, II, III, IVa, IVb, MB
	pk_1P2	MB
Inorganic pyrophosphatase (PPA)	*ppa_2P1*	IVa, MB
	ppa_3P1	MB
	ppa_4P1.6	III
PEP carboxylase (PEPC)	*ppc_1P3*	Ang, II, III, IVa, IVb, MB

	ppc_1P3_LGT:A	IVb
	ppc_1P3_LGT:M	Cim, III
	ppc_1P6	Cim, IVb
	ppc_1P7	II
Phosphoglycerate kinase (PGK)	-	no duplications
Pyruvate phosphate dikinase (PPDK)	*ppdk_1P1*	I, II, IVb, MB
	ppdk_1P2	MB
Pyruvate phosphate dikinase regulatory protein (PPDK-RP)	-	no duplications
PEP-phosphate translocator (PPT)	*ppt_1P4*	Ang, III, IVa
	ppt_1P6	II, III, IVa, IVb, MB
Sodium bile acid symporter 1 (SBAS)	-	no duplications
Tonoplast malate/fumarate transporter (TDT)	*tdt_1P2*	Ang, I, IVa, IVb, MB
Triosephosphate-phosphate translocator (TPT)	*tpt_1P2*	Cim

[1] I, II, III and IV refers to the nuclear clades of *A. semialata* (Chapter 4); Ang is *A. angusta*; Cim is *A. cimicina*; MB refers to *A. semialata* individuals with mixed genetic background (Chapter 4).

Table 3.S3. Read depth of transcriptome and genome data for polymorphic sites of *ppc* and *pck* genes of accessions of the genus *Alloteropsis*.

						Transcriptome			Low-coverage genome				High-coverage genome			
gene	**accession**	**site**	**exon**	**variant 1 ('major')**	**variant 2 ('minor')**	**dataset**	**depth major**	**depth minor**	**ratio**	**dataset**	**depth major**	**depth minor**	**ratio**	**dataset**	**depth major**	**depth minor**	**ratio**
pck1P1	BUR1	359	6	A	G	BF3	76	68	0.472	BUR1	0	0	-				
pck1P1	BUR1	548	7	C	T	BF3	195	81	0.293	BUR1	1	0	0.000				
pck1P1	BUR1	578	7	T	G	BF3	211	115	0.353	BUR1	1	0	0.000				
pck1P1	BUR1	617	8	T	G	BF3	179	118	0.397	BUR1	1	1	0.500				
pck1P1	BUR1	713	8	C	T	BF3	170	128	0.430	BUR1	2	0	0.000				
pck1P1	BUR1	1166	10	T	C	BF3	133	89	0.401	BUR1	1	0	0.000				
pck1P1	BUR1	1264	10	A	G	BF3	112	110	0.495	BUR1	0	0	-				
pck1P1	BUR1	1265	10	C	T	BF3	114	105	0.479	BUR1	0	0	-				
pck1P1	TAN1	1129	10	C	G	LO4	87	73	0.456	TAN1	2	0	0.000				
pck1P1	TAN1	1207	10	G	C	LO4	71	52	0.423	TAN1	0	0	-				
pck1P1	TAN1	1222	10	G	A	LO4	60	43	0.417	TAN1	0	0	-				
pck1P1	TAN1	1230	10	A	G	LO4	54	38	0.413	TAN1	0	0	-				
pck1P1	TAN2	1096	10	G	C	LO1	38	31	0.449	TAN2-A	0	2	1.000	TAN2-A	3	1	0.250
pck1P1	TAN2	1174	10	C	G	LO1	49	29	0.372	TAN2-A	0	1	1.000	TAN2-A	1	5	0.833
pck1P1	TAN2	1191	10	A	G	LO1	39	40	0.506	TAN2-A	0	1	1.000	TAN2-A	1	5	0.833
pck1P1	TAN2	1199	10	G	A	LO1	32	21	0.396	TAN2-A	0	1	1.000	TAN2-A	1	5	0.833
pck1P1	TAN4	1108	10	C	G	LO2	20	14	0.412	TAN4	1	2	0.667				
pck1P1	TAN4	1162	10	T	C	LO2	26	16	0.381	TAN4	1	0	0.000				
pck1P1	TAN4	1183	10	C	A	LO2	23	15	0.395	TAN4	1	0	0.000				
pck1P1	TAN4	1186	10	G	C	LO2	24	16	0.400	TAN4	1	0	0.000				

pck1P1	TAN4	1209	10	A	G	LO2	20	12	0.375	TAN4	0	0	-			
pck1P1LGT	AUS1	465	5	G	T	AUS2	5625	1877	0.250	AUS1	3	1	0.250			
pck1P1LGT	AUS1	848	8	G	A	AUS2	5446	1980	0.267	AUS1	6	0	0.000			
pck1P1LGT	AUS1	1165	8	C	T	AUS2	11537	3077	0.211	AUS1	2	0	0.000			
pck1P1LGT	AUS1	1242	9	C	T	AUS2	15476	2756	0.151	AUS1	2	0	0.000			
pck1P1LGT	AUS1	1453	10	G	A	AUS2	13538	1575	0.104	AUS1	7	0	0.000			
pck1P1LGT	BUR1	629	6	C	T	BF3	5109	721	0.124	BUR1	4	0	0.000			
pck1P1LGT	BUR1	656	6	C	G	BF3	5228	807	0.134	BUR1	3	0	0.000			
pck1P1LGT	BUR1	1021	8	C	T	BF3	6234	1159	0.157	BUR1	7	0	0.000			
pck1P1LGT	BUR1	1547	9	A	G	BF3	12754	3608	0.221	BUR1	3	0	0.000			
pck1P1LGT	BUR1	1669	10	G	A	BF3	6490	4378	0.403	BUR1	7	1	0.125			
pck1P1LGT	BUR1	1810	10	C	T	BF3	4192	3547	0.458	BUR1	4	1	0.200			
pck1P1LGT	MAD1	274	5	A	G	MAJ	78	57	0.422	MAD1	3	2	0.400			
pck1P1LGT	MAD1	1100	9	G	C	MAJ	477	83	0.148	MAD1	4	0	0.000			
pck1P1LGT	MAD1	1182	9	C	T	MAJ	825	63	0.071	MAD1	5	0	0.000			
pck1P1LGT	MAD1	1203	9	C	T	MAJ	691	47	0.064	MAD1	5	0	0.000			
pck1P1LGT	MAD1	1224	9	G	A	MAJ	611	42	0.064	MAD1	6	0	0.000			
pck1P1LGT	TAN4	1014	8	T	A	LO2	754	119	0.136	TAN4	0	3	1.000			
pck1P1LGT	TAN4	1134	9	G	C	LO2	1146	179	0.135	TAN4	0	3	1.000			
pck1P1LGT	TAN4	1211	9	C	G	LO2	1011	231	0.186	TAN4	0	4	1.000			
pck1P1LGT	TAN4	1236	9	G	A	LO2	960	199	0.172	TAN4	0	4	1.000			
pck1P1LGT	TAN4	1256	9	G	A	LO2	980	205	0.173	TAN4	0	4	1.000			
pck1P1LGT	TAN4	1295	10	C	T	LO2	939	146	0.135	TAN4	2	4	0.667			
pck1P1LGT	TPE1	299	5	C	T	TW10	1234	496	0.287	TPE1-3	5	1	0.167	TPE1-10	164	54	0.248
pck1P1LGT	TPE1	483	5	C	T	TW10	1389	602	0.302	TPE1-3	1	2	0.667	TPE1-10	105	53	0.335

pck1P1LGT	TPE1	783	8	C	T	TW10	1707	1032	0.377	TPE1-3	5	1	0.167	TPE1-10	99	50	0.336
pck1P1LGT	TPE1	1343	10	G	A	TW10	2664	364	0.120	TPE1-3	1	1	0.500	TPE1-10	91	19	0.173
ppc1P3	BUR1	267	2	C	T	BF3	2393	1802	0.430	BUR1	12	5	0.294				
ppc1P3	BUR1	268	2	G	A	BF3	3780	319	0.078	BUR1	10	7	0.412				
ppc1P3	BUR1	294	2	C	T	BF3	3660	326	0.082	BUR1	10	7	0.412				
ppc1P3	BUR1	295	2	G	A	BF3	3601	357	0.090	BUR1	15	2	0.118				
ppc1P3	BUR1	296	2	C	T	BF3	2370	1641	0.409	BUR1	16	1	0.059				
ppc1P3	BUR1	312	2	G	A	BF3	3131	1077	0.256	BUR1	9	5	0.357				
ppc1P3	BUR1	333	2	C	T	BF3	2039	1851	0.476	BUR1	10	4	0.286				
ppc1P3	BUR1	339	2	G	C	BF3	2197	1696	0.436	BUR1	10	5	0.333				
ppc1P3	BUR1	351	2	G	C	BF3	3586	384	0.097	BUR1	10	1	0.091				
ppc1P3	BUR1	355	2	C	T	BF3	3604	389	0.097	BUR1	10	2	0.167				
ppc1P3	BUR1	360	2	G	A	BF3	2078	1836	0.469	BUR1	5	4	0.444				
ppc1P3	BUR1	372	2	G	A	BF3	2497	2145	0.462	BUR1	7	3	0.300				
ppc1P3	BUR1	387	2	T	C	BF3	2540	2255	0.470	BUR1	5	5	0.500				
ppc1P3	BUR1	412	2	A	C	BF3	2457	2456	0.500	BUR1	9	7	0.438				
ppc1P3	BUR1	478	2	C	T	BF3	2576	2059	0.444	BUR1	14	4	0.222				
ppc1P3	BUR1	496	2	C	T	BF3	2519	1990	0.441	BUR1	8	12	0.600				
ppc1P3	BUR1	521	2	C	G	BF3	4736	508	0.097	BUR1	16	4	0.200				
ppc1P3	BUR1	537	2	G	T	BF3	4523	889	0.164	BUR1	16	1	0.059				
ppc1P3	BUR1	557	2	A	G	BF3	2993	2711	0.475	BUR1	8	7	0.467				
ppc1P3	BUR1	572	2	C	T	BF3	5318	494	0.085	BUR1	10	4	0.286				
ppc1P3	BUR1	575	2	C	T	BF3	3381	2605	0.435	BUR1	10	3	0.231				
ppc1P3	BUR1	601	2	T	C	BF3	3507	2990	0.460	BUR1	4	5	0.556				

ppc1P3	BUR1	631	2	C	T	BF3	3107	2650	0.460	BUR1	10	2	0.167
ppc1P3	BUR1	652	2	G	A	BF3	2847	2177	0.433	BUR1	5	4	0.444
ppc1P3	BUR1	703	3	G	A	BF3	3044	2145	0.413	BUR1	6	3	0.333
ppc1P3	BUR1	774	3	G	C	BF3	6193	1232	0.166	BUR1	11	0	0.000
ppc1P3	BUR1	823	4	G	A	BF3	3345	3229	0.491	BUR1	4	2	0.333
ppc1P3	BUR1	827	4	T	A	BF3	3340	3276	0.495	BUR1	6	2	0.250
ppc1P3	BUR1	875	4	T	A	BF3	2624	2437	0.482	BUR1	8	2	0.200
ppc1P3	BUR1	878	4	G	C	BF3	2727	2571	0.485	BUR1	5	4	0.444
ppc1P3	BUR1	891	4	T	C	BF3	4781	123	0.025	BUR1	11	2	0.154
ppc1P3	BUR1	900	4	A	T	BF3	4872	117	0.023	BUR1	9	2	0.182
ppc1P3	BUR1	924	4	G	A	BF3	2629	2221	0.458	BUR1	10	2	0.167
ppc1P3	BUR1	956	4	G	A	BF3	4945	162	0.032	BUR1	8	5	0.385
ppc1P3	BUR1	968	4	G	A	BF3	6432	140	0.021	BUR1	10	4	0.286
ppc1P3	BUR1	971	4	C	A	BF3	3619	2984	0.452	BUR1	11	4	0.267
ppc1P3	BUR1	972	4	G	A	BF3	5385	1256	0.189	BUR1	15	0	0.000
ppc1P3	BUR1	980	4	C	T	BF3	3370	3151	0.483	BUR1	6	8	0.571
ppc1P3	BUR1	983	4	C	T	BF3	3416	3212	0.485	BUR1	10	4	0.286
ppc1P3	BUR1	1036	4	T	C	BF3	4178	4007	0.490	BUR1	6	2	0.250
ppc1P3	BUR1	1183	6	C	T	BF3	4058	3590	0.469	BUR1	3	10	0.769
ppc1P3	BUR1	1201	6	T	G	BF3	4245	3952	0.482	BUR1	1	9	0.900
ppc1P3	BUR1	1218	6	C	T	BF3	4857	4633	0.488	BUR1	14	1	0.067
ppc1P3	BUR1	1246	6	T	C	BF3	5501	4399	0.444	BUR1	10	1	0.091
ppc1P3	BUR1	1295	7	T	C	BF3	5060	3799	0.429	BUR1	7	4	0.364
ppc1P3	BUR1	1310	7	C	G	BF3	4072	3489	0.461	BUR1	6	6	0.500

ppc1P3	BUR1	1320	7	T	C	BF3	3964	3439	0.465	BUR1	5	7	0.583
ppc1P3	BUR1	1326	7	G	A	BF3	3693	2967	0.445	BUR1	7	3	0.300
ppc1P3	BUR1	1336	7	C	T	BF3	3635	3558	0.495	BUR1	5	5	0.500
ppc1P3	BUR1	1353	7	G	A	BF3	7057	1042	0.129	BUR1	11	0	0.000
ppc1P3	BUR1	1378	7	C	T	BF3	4089	3675	0.473	BUR1	4	6	0.600
ppc1P3	BUR1	1399	7	A	G	BF3	4462	4378	0.495	BUR1	1	10	0.909
ppc1P3	BUR1	1467	8	T	C	BF3	3190	2843	0.471	BUR1	6	3	0.333
ppc1P3	BUR1	1496	8	C	T	BF3	3992	1205	0.232	BUR1	6	1	0.143
ppc1P3	BUR1	1513	8	G	T	BF3	3312	2781	0.456	BUR1	5	1	0.167
ppc1P3	BUR1	1525	8	C	G	BF3	3333	3000	0.474	BUR1	6	1	0.143
ppc1P3	BUR1	1590	8	G	A	BF3	5021	4083	0.448	BUR1	10	2	0.167
ppc1P3	BUR1	1631	8	G	C	BF3	4014	3292	0.451	BUR1	9	2	0.182
ppc1P3	BUR1	1637	8	T	C	BF3	3872	3270	0.458	BUR1	10	2	0.167
ppc1P3	BUR1	1651	8	C	G	BF3	3368	2724	0.447	BUR1	8	2	0.200
ppc1P3	BUR1	1667	8	G	A	BF3	3189	2480	0.437	BUR1	11	3	0.214
ppc1P3	BUR1	1672	8	A	G	BF3	4732	209	0.042	BUR1	11	3	0.214
ppc1P3	BUR1	1696	8	T	C	BF3	2650	2599	0.495	BUR1	8	4	0.333
ppc1P3	BUR1	1960	8	G	A	BF3	8598	840	0.089	BUR1	6	8	0.571
ppc1P3	BUR1	1970	8	G	A	BF3	9445	254	0.026	BUR1	8	4	0.333
ppc1P3	BUR1	2110	8	C	A	BF3	10285	585	0.054	BUR1	17	5	0.227
ppc1P3	BUR1	2758	9	C	T	BF3	6485	5979	0.480	BUR1	8	2	0.200
ppc1P3	BUR1	2766	9	C	A	BF3	6428	6368	0.498	BUR1	9	2	0.182
ppc1P3	BUR1	2788	9	C	G	BF3	7231	7147	0.497	BUR1	2	12	0.857
ppc1P3	BUR1	2809	9	G	A	BF3	7289	6976	0.489	BUR1	3	14	0.824

ppc1P3	BUR1	2831	9	G	C	BF3	6952	6952	0.500	BUR1	4	11	0.733
ppc1P3	BUR1	3137	9	C	G	BF3	12427	12325	0.498	BUR1	8	0	0.000
ppc1P3	BUR1	3463	9	G	T	BF3	11187	9668	0.464	BUR1	0	4	1.000
ppc1P3	BUR1	3518	10	G	T	BF3	13829	13501	0.494	BUR1	0	8	1.000
ppc1P3	BUR1	3535	10	T	C	BF3	13462	13082	0.493	BUR1	13	0	0.000
ppc1P3	BUR1	3802	10	C	G	BF3	7050	6522	0.481	BUR1	8	3	0.273
ppc1P3	BUR1	4047	10	A	G	BF3	5046	4653	0.480	BUR1	5	0	0.000
ppc1P3	MAD1	176	2	C	T	MAJ	80	3	0.036	MAD1	7	3	0.300
ppc1P3	MAD1	200	2	C	T	MAJ	37	23	0.383	MAD1	7	2	0.222
ppc1P3	MAD1	210	2	G	A	MAJ	56	4	0.067	MAD1	5	4	0.444
ppc1P3	MAD1	229	2	C	T	MAJ	33	2	0.057	MAD1	5	3	0.375
ppc1P3	MAD1	231	2	C	G	MAJ	30	5	0.143	MAD1	4	4	0.500
ppc1P3	MAD1	260	2	C	T	MAJ	30	6	0.167	MAD1	1	6	0.857
ppc1P3	MAD1	387	2	G	A	MAJ	59	49	0.454	MAD1	6	2	0.250
ppc1P3	MAD1	405	2	G	A	MAJ	106	10	0.086	MAD1	8	0	0.000
ppc1P3	MAD1	405	2	G	A	MAJ	87	9	0.094	MAD1	7	0	0.000
ppc1P3	MAD1	418	2	G	A	MAJ	80	15	0.158	MAD1	5	2	0.286
ppc1P3	MAD1	429	2	C	T	MAJ	72	15	0.172	MAD1	6	2	0.250
ppc1P3	MAD1	432	2	T	C	MAJ	84	38	0.311	MAD1	2	3	0.600
ppc1P3	MAD1	456	2	T	C	MAJ	94	29	0.236	MAD1	3	4	0.571
ppc1P3	MAD1	492	2	C	T	MAJ	88	26	0.228	MAD1	7	1	0.125
ppc1P3	MAD1	504	2	G	A	MAJ	92	10	0.098	MAD1	5	3	0.375
ppc1P3	MAD1	562	3	A	G	MAJ	89	50	0.360	MAD1	3	1	0.250
ppc1P3	MAD1	578	3	G	A	MAJ	108	33	0.234	MAD1	11	0	0.000

ppc1P3	MAD1	643	4	C	T	MAJ	178	30	0.144	MAD1	5	2	0.286
ppc1P3	MAD1	646	4	A	G	MAJ	143	56	0.281	MAD1	6	1	0.143
ppc1P3	MAD1	649	4	A	T	MAJ	145	55	0.275	MAD1	5	1	0.167
ppc1P3	MAD1	694	4	A	T	MAJ	117	53	0.312	MAD1	0	4	1.000
ppc1P3	MAD1	697	4	C	G	MAJ	166	38	0.186	MAD1	2	3	0.600
ppc1P3	MAD1	709	4	T	C	MAJ	169	45	0.210	MAD1	7	2	0.222
ppc1P3	MAD1	742	4	G	A	MAJ	261	22	0.078	MAD1	4	6	0.600
ppc1P3	MAD1	769	4	G	A	MAJ	265	84	0.241	MAD1	8	3	0.273
ppc1P3	MAD1	781	4	G	A	MAJ	311	74	0.192	MAD1	8	3	0.273
ppc1P3	MAD1	784	4	A	C	MAJ	207	187	0.475	MAD1	2	9	0.818
ppc1P3	MAD1	793	4	T	C	MAJ	227	132	0.368	MAD1	7	6	0.462
ppc1P3	MAD1	796	4	C	T	MAJ	224	200	0.472	MAD1	7	4	0.364
ppc1P3	MAD1	852	4	T	C	MAJ	241	218	0.475	MAD1	3	2	0.400
ppc1P3	MAD1	986	6	T	C	MAJ	276	270	0.495	MAD1	8	4	0.333
ppc1P3	MAD1	1004	6	G	T	MAJ	272	265	0.493	MAD1	6	2	0.250
ppc1P3	MAD1	1035	6	G	T	MAJ	447	81	0.153	MAD1	9	0	0.000
ppc1P3	MAD1	1046	6	T	C	MAJ	288	226	0.440	MAD1	5	2	0.286
ppc1P3	MAD1	1076	7	T	C	MAJ	246	220	0.472	MAD1	10	2	0.167
ppc1P3	MAD1	1091	7	G	C	MAJ	256	83	0.245	MAD1	2	8	0.800
ppc1P3	MAD1	1154	7	C	T	MAJ	270	77	0.222	MAD1	2	10	0.833
ppc1P3	MAD1	1175	7	A	G	MAJ	236	144	0.379	MAD1	0	11	1.000
ppc1P3	MAD1	1239	8	C	T	MAJ	213	50	0.190	MAD1	7	5	0.417
ppc1P3	MAD1	1283	8	T	G	MAJ	180	41	0.186	MAD1	5	7	0.583
ppc1P3	MAD1	1295	8	G	C	MAJ	182	20	0.099	MAD1	5	6	0.545

ppc1P3	MAD1	1320	8	C	A	MAJ	212	13	0.058	MAD1	5	2	0.286
ppc1P3	MAD1	1359	8	T	C	MAJ	178	31	0.148	MAD1	7	7	0.500
ppc1P3	MAD1	1446	8	C	T	MAJ	175	32	0.155	MAD1	7	6	0.462
ppc1P3	MAD1	1455	8	G	C	MAJ	204	13	0.060	MAD1	8	2	0.200
ppc1P3	MAD1	1498	8	T	C	MAJ	278	78	0.219	MAD1	3	9	0.750
ppc1P3	MAD1	2170	8	T	C	MAJ	511	128	0.200	MAD1	4	6	0.600
ppc1P3	MAD1	2237	9	T	A	MAJ	835	157	0.158	MAD1	4	4	0.500
ppc1P3	MAD1	2347	9	T	G	MAJ	1444	273	0.159	MAD1	8	5	0.385
ppc1P3	MAD1	2353	9	G	A	MAJ	1444	303	0.173	MAD1	8	5	0.385
ppc1P3	MAD1	2375	9	G	C	MAJ	1458	330	0.185	MAD1	9	5	0.357
ppc1P3	MAD1	2388	9	T	G	MAJ	1846	112	0.057	MAD1	10	2	0.167
ppc1P3	MAD1	2574	9	G	C	MAJ	3433	523	0.132	MAD1	4	7	0.636
ppc1P3	MAD1	2673	9	G	T	MAJ	4078	387	0.087	MAD1	2	5	0.714
ppc1P3	MAD1	2702	10	G	T	MAJ	4635	410	0.081	MAD1	5	6	0.545
ppc1P3	MAD1	2712	10	C	T	MAJ	3983	436	0.099	MAD1	5	5	0.500
ppc1P3	MAD1	2754	10	C	T	MAJ	3281	446	0.120	MAD1	4	5	0.556
ppc1P3	MAD1	3002	10	G	A	MAJ	1836	123	0.063	MAD1	4	2	0.333
ppc1P3	RSA2	300	2	G	C	KWT3	85	55	0.393	RSA2	4	0	0.000
ppc1P3	RSA2	366	2	A	C	KWT3	63	50	0.442	RSA2	2	1	0.333
ppc1P3	RSA2	405	2	G	C	KWT3	69	49	0.415	RSA2	2	2	0.500
ppc1P3	RSA2	609	2	A	C	KWT3	97	86	0.470	RSA2	2	1	0.333
ppc1P3	RSA2	627	3	G	T	KWT3	94	91	0.492	RSA2	2	1	0.333
ppc1P3	RSA2	636	3	G	C	KWT3	101	95	0.485	RSA2	2	2	0.500
ppc1P3	RSA2	669	3	A	G	KWT3	113	88	0.438	RSA2	3	2	0.400

ppc1P3	RSA2	690	3	A	G	KWT3	120	96	0.444	RSA2	0	2	1.000
ppc1P3	RSA2	750	4	A	G	KWT3	100	96	0.490	RSA2	0	1	1.000
ppc1P3	RSA2	762	4	T	A	KWT3	90	77	0.461	RSA2	1	0	0.000
ppc1P3	RSA2	837	4	G	A	KWT3	105	81	0.435	RSA2	4	0	0.000
ppc1P3	RSA2	849	4	G	A	KWT3	116	103	0.470	RSA2	5	0	0.000
ppc1P3	RSA2	861	4	T	C	KWT3	109	107	0.495	RSA2	6	0	0.000
ppc1P3	RSA2	1027	6	C	T	KWT3	155	84	0.351	RSA2	2	0	0.000
ppc1P3	RSA2	1048	6	C	T	KWT3	164	92	0.359	RSA2	2	0	0.000
ppc1P3	RSA2	1066	6	T	G	KWT3	171	89	0.342	RSA2	2	0	0.000
ppc1P3	RSA2	1108	6	T	C	KWT3	147	114	0.437	RSA2	2	0	0.000
ppc1P3	RSA2	1144	7	T	G	KWT3	144	105	0.422	RSA2	0	0	-
ppc1P3	RSA2	1162	7	C	T	KWT3	131	116	0.470	RSA2	1	0	0.000
ppc1P3	RSA2	1213	7	G	T	KWT3	144	115	0.444	RSA2	3	0	0.000
ppc1P3	RSA2	1222	7	A	G	KWT3	137	112	0.450	RSA2	3	0	0.000
ppc1P3	RSA2	1225	7	T	C	KWT3	137	112	0.450	RSA2	3	0	0.000
ppc1P3	RSA2	1303	8	G	A	KWT3	123	115	0.483	RSA2	0	0	-
ppc1P3	RSA2	1342	8	G	A	KWT3	129	101	0.439	RSA2	0	2	1.000
ppc1P3	RSA2	1462	8	T	G	KWT3	134	101	0.430	RSA2	3	0	0.000
ppc1P3	RSA2	1465	8	C	G	KWT3	135	100	0.426	RSA2	3	0	0.000
ppc1P3	RSA2	1516	8	G	C	KWT3	106	105	0.498	RSA2	0	1	1.000
ppc1P3	RSA2	1543	8	C	G	KWT3	170	114	0.401	RSA2	2	0	0.000
ppc1P3	RSA2	1738	8	G	C	KWT3	151	149	0.497	RSA2	0	3	1.000
ppc1P3	RSA2	1766	8	T	G	KWT3	140	102	0.421	RSA2	3	0	0.000
ppc1P3	RSA2	1777	8	T	C	KWT3	134	83	0.382	RSA2	0	3	1.000

ppc1P3	RSA2	1779	8	G	C	KWT3	138	78	0.361	RSA2	3	0	0.000			
ppc1P3	RSA2	1785	8	G	C	KWT3	135	69	0.338	RSA2	2	0	0.000			
ppc1P3	RSA2	1806	8	C	G	KWT3	137	81	0.372	RSA2	4	0	0.000			
ppc1P3	RSA2	1872	8	C	G	KWT3	144	140	0.493	RSA2	3	0	0.000			
ppc1P3	RSA2	1919	8	G	T	KWT3	138	104	0.430	RSA2	2	0	0.000			
ppc1P3	RSA2	1963	8	C	T	KWT3	117	85	0.421	RSA2	3	0	0.000			
ppc1P3	RSA2	1991	8	C	G	KWT3	123	101	0.451	RSA2	4	0	0.000			
ppc1P3	RSA2	2048	8	G	C	KWT3	127	116	0.477	RSA2	4	0	0.000			
ppc1P3	RSA2	2084	8	G	C	KWT3	122	115	0.485	RSA2	3	0	0.000			
ppc1P3	RSA2	2161	8	T	C	KWT3	163	137	0.457	RSA2	0	2	1.000			
ppc1P3	RSA2	2181	8	A	G	KWT3	179	147	0.451	RSA2	0	1	1.000			
ppc1P3	RSA2	2182	8	A	C	KWT3	179	147	0.451	RSA2	0	1	1.000			
ppc1P3	RSA2	2546	9	C	T	KWT3	324	275	0.459	RSA2	2	0	0.000			
ppc1P3	RSA2	2625	9	G	C	KWT3	318	281	0.469	RSA2	3	0	0.000			
ppc1P3	RSA2	2631	9	C	T	KWT3	323	284	0.468	RSA2	3	0	0.000			
ppc1P3	RSA2	2647	9	A	T	KWT3	288	251	0.466	RSA2	4	0	0.000			
ppc1P3	RSA2	2953	10	C	G	KWT3	391	268	0.407	RSA2	1	0	0.000			
ppc1P3	TAN1	302	2	C	G	LO4	258	92	0.263	TAN1	0	9	1.000			
ppc1P3	TAN1	369	2	C	A	LO4	309	105	0.254	TAN1	0	9	1.000			
ppc1P3	TAN1	859	4	G	A	LO4	834	440	0.345	TAN1	10	0	0.000			
ppc1P3	TAN1	2458	9	A	C	LO4	1312	86	0.062	TAN1	7	0	0.000			
ppc1P3	TAN2	306	2	C	G	LO1	253	118	0.318	TAN2-A	1	1	0.500	TAN2-A	8	8	0.500
ppc1P3	TAN2	373	2	C	A	LO1	301	186	0.382	TAN2-A	0	2	1.000	TAN2-A	10	8	0.444
ppc1P3	TAN2	731	4	T	C	LO1	2067	341	0.142	TAN2-A	5	1	0.167	TAN2-A	17	4	0.190

ppc1P3	TAN2	870	4	G	A	LO1	1483	570	0.278	TAN2-A	3	2	0.400	TAN2-A	20	5	0.200
ppc1P3	TAN4	250	2	T	C	LO2	46	3	0.061	TAN4	3	5	0.625				
ppc1P3	TAN4	262	2	G	C	LO2	38	7	0.156	TAN4	10	0	0.000				
ppc1P3	TAN4	274	2	T	C	LO2	46	3	0.061	TAN4	5	6	0.545				
ppc1P3	TAN4	287	2	C	T	LO2	42	3	0.067	TAN4	5	4	0.444				
ppc1P3	TAN4	294	2	G	T	LO2	41	3	0.068	TAN4	5	4	0.444				
ppc1P3	TAN4	299	2	A	T	LO2	41	3	0.068	TAN4	4	0	0.000				
ppc1P3	TAN4	321	2	G	A	LO2	25	15	0.375	TAN4	7	0	0.000				
ppc1P3	TAN4	327	2	A	G	LO2	42	4	0.087	TAN4	3	3	0.500				
ppc1P3	TAN4	648	3	A	T	LO2	85	15	0.150	TAN4	5	0	0.000				
ppc1P3	TAN4	1507	8	C	G	LO2	27	16	0.372	TAN4	0	5	1.000				
ppc1P3	TAN4	1553	8	T	C	LO2	23	21	0.477	TAN4	0	7	1.000				
ppc1P3	TAN4	1631	8	C	G	LO2	25	4	0.138	TAN4	1	5	0.833				
ppc1P3	TAN4	1651	8	C	T	LO2	23	5	0.179	TAN4	4	2	0.333				
ppc1P3	TAN4	1693	8	G	C	LO2	25	19	0.432	TAN4	3	3	0.500				
ppc1P3	TAN4	1740	8	G	A	LO2	45	13	0.224	TAN4	6	4	0.400				
ppc1P3	TAN4	1754	8	A	G	LO2	39	13	0.250	TAN4	6	2	0.250				
ppc1P3	TAN4	1867	8	G	C	LO2	22	19	0.463	TAN4	2	3	0.600				
ppc1P3	TAN4	2125	8	G	C	LO2	30	7	0.189	TAN4	2	0	0.000				
ppc1P3	TAN4	2140	8	G	A	LO2	30	7	0.189	TAN4	2	0	0.000				
ppc1P3	TAN4	2149	8	G	C	LO2	31	6	0.162	TAN4	5	0	0.000				
ppc1P3	TAN4	2151	8	A	C	LO2	31	8	0.205	TAN4	4	2	0.333				
ppc1P3	TAN4	2556	9	C	G	LO2	86	72	0.456	TAN4	0	4	1.000				
ppc1P3	TPE1	292	2	G	A	TW10	1478	385	0.207	TPE1-3	2	5	0.714	TPE1-10	141	69	0.329
ppc1P3	TPE1	420	2	G	A	TW10	1379	527	0.276	TPE1-3	10	0	0.000	TPE1-10	149	59	0.284

ppc1P3	TPE1	587	2	C	A	TW10	1697	572	0.252	TPE1-3	5	0	0.000	TPE1-10	148	49	0.249
ppc1P3	TPE1	590	2	G	A	TW10	1701	575	0.253	TPE1-3	3	2	0.400	TPE1-10	144	50	0.258
ppc1P3	TPE1	645	3	G	A	TW10	1168	910	0.438	TPE1-3	3	0	0.000	TPE1-10	72	66	0.478
ppc1P3	TPE1	739	4	A	T	TW10	2203	786	0.263	TPE1-3	0	10	1.000	TPE1-10	111	30	0.213
ppc1P3	TPE1	1332	7	C	T	TW10	2093	1088	0.342	TPE1-3	5	0	0.000	TPE1-10	158	61	0.279
ppc1P3	TPE1	1695	8	T	C	TW10	2729	1284	0.320	TPE1-3	0	7	1.000	TPE1-10	109	39	0.264
ppc1P3LGTA	AUS1	849	4	A	T	AUS2	1426	396	0.217	AUS1	5	0	0.000				
ppc1P3LGTA	AUS1	1136	7	A	T	AUS2	1678	1592	0.487	AUS1	1	1	0.500				
ppc1P3LGTA	AUS1	1213	7	C	A	AUS2	1793	1662	0.481	AUS1	0	1	1.000				
ppc1P3LGTA	AUS1	3095	10	G	T	AUS2	4540	1437	0.240	AUS1	2	0	0.000				
ppc1P3LGTM	BUR1	1238	6	C	T	BF3	4414	1303	0.228	BUR1	3	0	0.000				
ppc1P3LGTM	BUR1	1350	7	T	G	BF3	3948	1350	0.255	BUR1	3	0	0.000				
ppc1P3LGTM	BUR1	2194	8	T	G	BF3	2826	530	0.158	BUR1	1	0	0.000				
ppc1P3LGTM	BUR1	2211	8	C	G	BF3	1570	1546	0.496	BUR1	1	0	0.000				
ppc1P3LGTM	BUR1	2953	9	C	T	BF3	12262	3869	0.240	BUR1	2	1	0.333				
ppc1P3LGTM	MAD1	727	4	A	G	MAJ	78	30	0.278	MAD1	4	0	0.000				
ppc1P3LGTM	MAD1	838	4	C	T	MAJ	56	25	0.309	MAD1	0	2	1.000				
ppc1P3LGTM	MAD1	2202	8	A	G	MAJ	157	47	0.230	MAD1	4	0	0.000				
ppc1P3LGTM	MAD1	2244	8	C	T	MAJ	307	23	0.070	MAD1	1	0	0.000				
ppc1P3LGTM	MAD1	2268	9	G	T	MAJ	326	32	0.089	MAD1	0	0	-				
ppc1P3LGTM	MAD1	2269	9	G	T	MAJ	326	32	0.089	MAD1	0	0	-				
ppc1P6	BUR1	1099	6	G	A	BF3	622	571	0.479	BUR1	0	0	-				
ppc1P6	MAD1	2091	8	A	G	MAJ	106	10	0.086	MAD1	0	3	1.000				
ppc1P6	RSA2	873	4	G	A	KWT3	36	17	0.321	RSA2	0	1	1.000				
ppc1P6	RSA2	1115	7	C	T	KWT3	59	3	0.048	RSA2	0	0	-				

ppc1P6	RSA2	1126	7	T	G	KWT3	58	3	0.049	RSA2	0	0	-				
ppc1P6	RSA2	1189	7	G	A	KWT3	63	3	0.045	RSA2	0	0	-				
ppc1P6	RSA2	1338	8	C	T	KWT3	83	5	0.057	RSA2	0	0	-				
ppc1P6	RSA2	1491	8	T	C	KWT3	38	31	0.449	RSA2	0	0	-				
ppc1P6	RSA2	1496	8	G	A	KWT3	72	3	0.040	RSA2	0	0	-				
ppc1P6	RSA2	1611	8	A	G	KWT3	90	4	0.043	RSA2	0	0	-				
ppc1P6	RSA2	1669	8	C	G	KWT3	87	4	0.044	RSA2	0	0	-				
ppc1P6	RSA2	1691	8	G	A	KWT3	87	3	0.033	RSA2	0	0	-				
ppc1P6	RSA2	2207	8	A	G	KWT3	134	11	0.076	RSA2	0	0	-				
ppc1P6	RSA2	2227	8	T	C	KWT3	147	12	0.075	RSA2	0	0	-				
ppc1P6	RSA2	2680	10	C	T	KWT3	243	8	0.032	RSA2	0	0	-				
ppc1P6	RSA2	2695	10	T	C	KWT3	295	6	0.020	RSA2	0	0	-				
ppc1P6	RSA2	2725	10	T	G	KWT3	337	7	0.020	RSA2	0	0	-				
ppc1P6	TAN1	2179	8	A	G	LO4	1021	956	0.484	TAN1	0	0	-				
ppc1P6	TAN2	2190	8	A	G	LO1	1544	1522	0.496	TAN2-A	0	2	1.000	TAN2-A	8	11	0.579
ppc1P6	TPE1	234	2	T	A	TW10	4907	4832	0.496	TPE1-3	1	5	0.833	TPE1-10	80	87	0.521
ppc1P6	TPE1	292	2	A	T	TW10	6130	2198	0.264	TPE1-3	4	6	0.600	TPE1-10	149	47	0.240
ppc1P6	TPE1	343	2	T	C	TW10	7148	6748	0.486	TPE1-3	8	5	0.385	TPE1-10	105	100	0.488
ppc1P6	TPE1	545	2	G	A	TW10	9872	1166	0.106	TPE1-3	13	0	0.000	TPE1-10	194	30	0.134
ppc1P6	TPE1	590	2	A	G	TW10	8513	2790	0.247	TPE1-3	11	0	0.000	TPE1-10	170	51	0.231
ppc1P6	TPE1	635	2	A	C	TW10	9741	3208	0.248	TPE1-3	3	2	0.400	TPE1-10	165	48	0.225
ppc1P6	TPE1	684	3	A	G	TW10	9909	3924	0.284	TPE1-3	6	3	0.333	TPE1-10	166	51	0.235
ppc1P6	TPE1	746	4	G	A	TW10	10349	1430	0.121	TPE1-3	6	1	0.143	TPE1-10	175	21	0.107
ppc1P6	TPE1	773	4	A	G	TW10	8157	3079	0.274	TPE1-3	9	1	0.100	TPE1-10	162	45	0.217
ppc1P6	TPE1	774	4	C	G	TW10	9874	1443	0.128	TPE1-3	9	1	0.100	TPE1-10	184	24	0.115

ppc1P6	TPE1	832	4	G	T	TW10	9379	3148	0.251	TPE1-3	5	1	0.167	TPE1-10	151	52	0.256
ppc1P6	TPE1	935	4	T	C	TW10	12325	4028	0.246	TPE1-3	4	0	0.000	TPE1-10	169	33	0.163
ppc1P6	TPE1	1136	6	C	T	TW10	13550	1635	0.108	TPE1-3	4	0	0.000	TPE1-10	203	29	0.125
ppc1P6	TPE1	1334	7	A	G	TW10	12780	4294	0.251	TPE1-3	2	3	0.600	TPE1-10	166	62	0.272
ppc1P6	TPE1	1364	8	C	T	TW10	14543	4966	0.255	TPE1-3	0	1	1.000	TPE1-10	163	54	0.249
ppc1P6	TPE1	1400	8	C	T	TW10	13988	2063	0.129	TPE1-3	3	0	0.000	TPE1-10	191	36	0.159
ppc1P6	TPE1	1518	8	C	T	TW10	24632	3445	0.123	TPE1-3	14	0	0.000	TPE1-10	215	41	0.160
ppc1P6	TPE1	1594	8	A	G	TW10	10480	6114	0.368	TPE1-3	10	1	0.091	TPE1-10	136	80	0.370
ppc1P6	TPE1	1703	8	A	G	TW10	22019	2984	0.119	TPE1-3	11	1	0.083	TPE1-10	171	31	0.153
ppc1P6	TPE1	1759	8	G	A	TW10	22892	2874	0.112	TPE1-3	11	0	0.000	TPE1-10	181	17	0.086
ppc1P6	TPE1	1812	8	G	T	TW10	17538	2977	0.145	TPE1-3	11	3	0.214	TPE1-10	168	34	0.168
ppc1P6	TPE1	1905	8	C	T	TW10	25978	4227	0.140	TPE1-3	8	3	0.273	TPE1-10	187	37	0.165
ppc1P6	TPE1	2051	8	C	T	TW10	20246	3035	0.130	TPE1-3	9	1	0.100	TPE1-10	191	37	0.162
ppc1P6	TPE1	2078	8	A	G	TW10	22324	3090	0.122	TPE1-3	9	1	0.100	TPE1-10	184	30	0.140
ppc1P6	TPE1	2230	8	G	A	TW10	25374	3961	0.135	TPE1-3	4	4	0.500	TPE1-10	183	26	0.124
ppc1P6	TPE1	2804	9	C	T	TW10	17988	2513	0.123	TPE1-3	3	0	0.000	TPE1-10	184	22	0.107
ppc1P6	TPE1	2945	9	G	A	TW10	30514	4681	0.133	TPE1-3	6	2	0.250	TPE1-10	192	24	0.111
ppc1P6	TPE1	3055	9	G	A	TW10	29902	4754	0.137	TPE1-3	4	0	0.000	TPE1-10	205	33	0.139
ppc1P6	TPE1	3187	10	C	T	TW10	28165	16983	0.376	TPE1-3	6	1	0.143	TPE1-10	99	56	0.361
ppc1P6	TPE1	3369	10	G	A	TW10	25262	3876	0.133	TPE1-3	11	0	0.000	TPE1-10	129	24	0.157

Table 3.S4. Allele-specific expression analyses.

Gene	Accession	Type	Sequencing depth	Number of SNPs	Mean depth[1] (transcript)	Mean depth[1] (genome)	Regression analysis[2]		
							Slope	R^2	*p*-value
pck1P1_LGT	AUS1	C_4	low-coverage	5	12,577	4.2	0.27	0	0.46
pck1P1_LGT	BUR1	C_4	low-coverage	6	9,038	5	**1.6**	**0.91**	**< 0.001**
pck1P1_LGT	MAD1	C_4	low-coverage	5	595	5	**0.84**	**0.93**	**0.01**
pck1P1_LGT	TAN4	C_4	low-coverage	6	1,145	4	0.08	0	0.38
pck1P1_LGT	TPE1-3	C_4	low-coverage	4	2,372	4.2	-0.19	0	0.57
pck1P1	BUR1	C_4	low-coverage	5	284	1.4	0.06	0	0.71
pck1P1	TAN2-A	C_3+C_4	low-coverage	4	70	1.2	-	-	-
pck1P1	TAN4	C_4	low-coverage	4	38	1.5	0.03	0.41	0.22
ppc1P3_LGT_A	AUS1	C_4	low-coverage	4	3,631	2.5	0.28	0.7	0.1
ppc1P3_LGT_M	BUR1	C_4	low-coverage	5	6,724	2.2	-0.13	0	0.81
ppc1P3_LGT_M	MAD1	C_4	low-coverage	4	181	2.8	0.12	0	0.45
ppc1P3	BUR1	C_4	low-coverage	77	7,760	12.2	0.2	0.07	0.01
ppc1P3	MAD1	C_4	low-coverage	58	748	9.3	0.03	0	0.74
ppc1P3	RSA2	C_3	low-coverage	46	268	2.8	0.03	0.05	0.09
ppc1P3	TAN1	C_3+C_4	low-coverage	4	859	8.8	0.06	0	0.74
ppc1P3	TAN2-A	C_3+C_4	low-coverage	4	1,330	3.8	0.26	0.74	0.09
ppc1P3	TAN4	C_4	low-coverage	22	50	6.4	0.17	0.13	0.06
ppc1P3	TPE1-3	C_4	low-coverage	8	2,572	6.5	-0.06	0	0.38
ppc1P6	TPE1-3	C_4	low-coverage	30	20,032	8.4	**0.21**	**0.23**	**< 0.001**
pck1P1_LGT	TPE1-10	C_4	high-coverage	4	2,372	158.8	1.26	0.75	0.09
pck1P1	TAN2-A	C_3+C_4	high-coverage	4	70	5.5	-0.04	0	0.8
ppc1P3	TAN2-A	C_3+C_4	high-coverage	4	1,330	20	0.48	0.38	0.23
ppc1P3	TPE1-10	C_4	high-coverage	8	2,572	181.9	0.61	0.38	0.06
ppc1P6	TPE1-10	C_4	high-coverage	30	20,032	209.3	**0.99**	**0.93**	**< 0.001**

[1] Mean number of reads covering each SNP; [2] Linear regression of the depth of the minor allele in the transcriptome and genome datasets.

Table 3.S5. Effect of phylogenetic tree on the phylogenetic generalized least squares (PGLS) analysis used to test for an association between changes in gene copy number and changes in transcript abundance.

Gene family	*p*-value range[1]
Alanine aminotransferase (ALA-AT)	0.041 – 0.279
Aspartate aminotransferase (ASP-AT)	0.288 – 0.536
Carbonic anhydrase (CA)	0.392 – 0.62
Dicarboxylate transporter (DIT)	-
NAD-malate dehydrogenase (NAD-MDH)	0.061 – 0.224
NAD-malic enzyme (NAD-ME)	0.499 – 0.633
NADP-malate dehydrogenase (NADP-MDH)	-
NADP-malic enzyme (NADP-ME)	0.405 – 0.591
PEP carboxykinase (PCK)	0.001 – 0.006
PEP carboxylase (PEPC)	< 0.001
Pyruvate phosphate dikinase (PPDK)	0.798 – 0.835
PEP-phosphate translocator (PPT)	0.557 – 0.764
Sodium bile acid symporter (SBAS)	-
Triosephosphate-phosphate translocator (TPT)	-

[1] *p*-value ranges are the interquartile range of PGLS fitting computed using 100 independent Bayesian trees. Before the analysis, transcript abundance values were log10 transformed and copy numbers were expressed as integers. Gene families lacking *p*-values do not show copy number variation, or contain representatives with no gene sequence available for the phylogenetic analysis. *p*-value ranges in bold include statistically significant results after correcting the significance level ($\alpha = 0.05$) for multiple comparisons.

Chapter 4.

Genome biogeography reveals the intraspecific spread of adaptive mutations for a complex trait

Chapter 4. Genome biogeography reveals the intraspecific spread of adaptive mutations for a complex trait

Jill K. Olofsson[1*], Matheus Bianconi[1*], Guillaume Besnard[2], Luke T. Dunning[1], Marjorie R. Lundgren[1], Helene Holota[2], Maria S. Vorontsova[3], Oriane Hidalgo[3], Ilia J. Leitch[3], Patrik Nosil[1], Colin P. Osborne[1], Pascal-Antoine Christin[1]

[*] These authors contributed equally to this work

[1] Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK.

[2] Laboratoire Evolution & Diversité e Biologique (EDB UMR5174), Université de Toulouse, CNRS, ENSFEA, UPS, 118 route de Narbonne, F-31062, Toulouse, France.

[3] Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AB, UK.

This work was published in **Molecular Ecology**, Volume 25, Issue 24, 14 December 2016, Pages 6107-6123.

Personal contribution: I performed the analyses of the laterally-acquired genes. I co-wrote the manuscript with Dr. Jill K. Olofsson, who performed the genome-wide analyses. All co-authors commented on the text before submission.

4.1. Abstract

Physiological novelties are often studied at macro-evolutionary scales such that their micro-evolutionary origins remain poorly understood. Here, we test the hypothesis that key components of a complex trait can evolve in isolation and later be combined by gene flow. We use C_4 photosynthesis as a study system, a derived physiology that increases plant productivity in warm, dry conditions. The grass *Alloteropsis semialata* includes C_4 and non-C_4 genotypes, with some populations using laterally acquired C_4-adaptive loci, providing an outstanding system to track the spread of novel adaptive mutations. Using genome data from C_4 and non-C_4 *A. semialata* individuals spanning the species' range, we infer and date past migrations of different parts of the genome. Our results show that photosynthetic types initially diverged in isolated populations, where key C_4 components were acquired. However, rare but recurrent subsequent gene flow allowed the spread of adaptive loci across genetic pools. Indeed, laterally acquired genes for key C_4 functions were rapidly passed between populations with otherwise distinct genomic backgrounds. Thus, our intraspecific study of C_4-related genomic variation indicates that components of adaptive traits can evolve separately and later be combined through secondary gene flow, leading to the assembly and optimization of evolutionary innovations.

Keywords: adaptation, C_4 photosynthesis, gene flow, lateral gene transfer, phylogeography.

4.2. Introduction

Over evolutionary time, living organisms have been able to colonize almost every possible environment, often via the acquisition of novel adaptations. While impressive changes can be observed across phyla, adaptive evolution by natural selection occurs within populations (e.g. Geber and Griffen 2003; Morjan and Rieseberg 2004). For most complex adaptive novelties, the intraspecific dynamics that lead to their progressive emergence are poorly understood. Indeed, if novel complex traits gain their function only when multiple anatomical and/or biochemical components work together, the order of acquisition of such components raises intriguing questions (Meléndez-Hevia et al. 1996; Lenski et al. 2003). One possibility is that the acquisition of one key component is sufficient to trigger a novel trait (e.g. Ourisson and Nakatani 1994), allowing the subsequent selection of novel mutations for the other components in the genetic pool that fixed the first component. The alternative would assume that components accumulate independently of each other in isolated populations and are later assembled by secondary gene flow and subsequent selection to form the complex trait (Morjan and Rieseberg 2004; Leinonen et al. 2006; Hufford et al. 2013; Ellstrand 2014; Miller et al. 2014). Differentiating these scenarios requires the inference of the order of mutations for a novel complex trait, as well as their past spread throughout the history of divergence, migration and secondary gene flow in one or several related species. Such investigations must rely on study systems in which variation in an adaptive complex trait, and its underlying genomic basis, can be traced back through time.

C_4 photosynthesis is a physiological state, present in ~3% of plant species (Sage 2016), which results from the co-ordinated action of multiple enzymes and anatomical components (Hatch 1987; Christin and Osborne 2014). C_4 biochemistry relies on well-characterized enzymes that also exist in non-C_4 plants, but with altered abundance, cellular and subcellular localization, regulation and kinetics (Kanai and Edwards 1999). The main effect of C_4 photosynthesis is an increase in CO_2 concentration at the place of its fixation by the enzyme Rubisco in the Calvin–Benson cycle (von Caemmerer and Furbank 2003). This is advantageous in conditions that restrict CO_2 availability, especially in warm and arid environments under the low-CO_2 atmosphere that has prevailed for the last 30 million years (Sage et al. 2012). C_4 plants consequently dominate most open biomes in tropical and subtropical regions, where they achieve high growth rates and large biomass (Griffith et al. 2015; Atkinson et al. 2016). Despite its

apparent complexity, C_4 photosynthesis evolved more than 60 times independently over the ancestral C_3 type (Sage et al. 2011), and evolutionary transitions were facilitated by the existence of anatomical and genetic enablers in some groups of plants (Christin et al. 2013b, 2015). However, the micro-evolutionary history of photosynthetic transitions is yet to be addressed.

Most C_4 lineages evolved this photosynthetic system millions of years ago, so that the initial changes linked to C_4 evolution remain obscured (Christin and Osborne 2014). In a couple of groups, closely related species present a spectrum of more or less complete C_4 traits, which is interpreted as the footprint of the gradual evolution of C_4 (e.g. McKown et al. 2005; Christin et al. 2011; Fisher et al. 2015). These groups provide powerful systems to reconstruct the order of changes during the transition to C_4 photosynthesis (e.g. McKown and Dengler 2007; Heckmann et al. 2013; Williams et al. 2013). However, the presumed lack of gene flow among these related species impedes testing hypotheses about the importance of secondary gene flow mixing mutations that were fixed in isolated populations. So far, the presence of genotypes with different photosynthetic types has been reported in only one taxon, the grass *Alloteropsis semialata*.

Alloteropsis semialata includes C_3 and C_4 individuals (Ellis 1974), and a recent study further described individuals with only some of the C_4 anatomical and biochemical components, which allow a weak C_4 cycle (i.e. C_3–C_4 intermediates; Lundgren et al. 2016). Other species in this genus, *A. angusta*, *A. cimicina*, *A. paniculata* and *A. papillosa*, are C_4, but perform the C_4 cycle using different enzymes and leaf tissues than *A. semialata*, which points to independent realizations of the C_4 phenotype (Christin et al. 2010). Analyses of genes for key C_4 enzymes in a handful of accessions have revealed that some populations of *A. semialata* carry C_4 genes that have been laterally acquired from distant C_4 relatives (Christin et al. 2012). The laterally acquired genes include one for phosphoenolpyruvate carboxykinase (*pck*) and three different copies for phosphoenolpyruvate carboxylase (*ppc*). These laterally acquired genes are integrated into the C_4 cycle of some extant accessions of *Alloteropsis* (Christin et al. 2012, 2013a), but genes for other C_4 enzymes have been transmitted following the species tree (vertically inherited), and gained their C_4 function via novel mutations (Christin et al. 2013a). Some C_4 *Alloteropsis* populations presumably still use the vertically inherited *ppc* and *pck* homologs for their C_4 cycles. However, the laterally acquired *ppc* and *pck* copies spent millions of years in other C_4 species, where they

acquired adaptive mutations that likely increased their fit for the C_4 function before their transfer (Christin et al. 2012). The potential adaptive value of the laterally acquired genes, as well as their restriction to some C_4 populations, provides a tractable system to elucidate gene movements that led to the emergence and strengthening of the complex C_4 adaptive trait. However, the geographical distributions and frequencies of these laterally acquired genes are still poorly understood, and the genome history of *A. semialata* remains largely unexplored.

In this study, we obtain low-coverage whole-genome sequencing data from *A. semialata* individuals spread across the species' geographical range and differing in photosynthetic type. We use the data to first infer the history of isolation and secondary contact, and then to track the acquisition and spread of the laterally acquired genes. This biogeographic framework allows us to test whether the C_4 complex trait was assembled via the sequential fixation of novel mutations within each isolated gene pool or via gene flow combining mutations that had been fixed in distinct gene pools (Fig. 4.1). In the first scenario, the history of C_4-adaptive mutations, represented by the laterally acquired genes, would correspond to the sequence of migration and isolation of populations and largely match the history of the rest of the genome (Fig. 4.1A). In the second scenario, the history of C_4-adaptive mutations would differ from that of the rest of the genome, their selection-driven spread across genetic lineages resulting in more recent coalescence times and gene topologies that differ from the species topology (Fig. 4.1B). This first intraspecific spatial genomic analysis of key components of the C_4 complex trait opens new avenues to understand the micro-evolutionary processes that led to macro-evolutionary innovations.

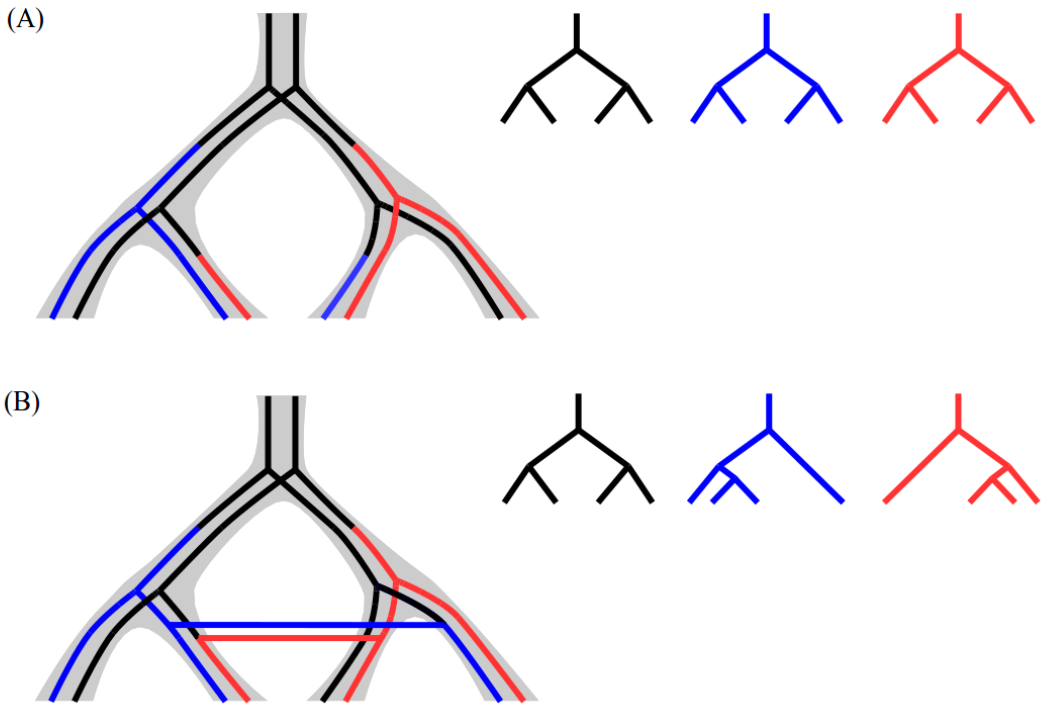


Fig. 4.1. Competing scenarios for the assembly of a complex trait. (**A)** The trait is assembled by sequential fixation mutations within each genetic pool; (**B)** Mutations that were fixed in isolation are later assembled via secondary gene flow. The species tree is outlined by thick grey branches, and colored branches indicate novel mutations on individual genes. Individual gene trees are drawn in blue and red on the right. In scenario (**A)** the histories of adaptive mutations correspond to the history of the rest of the genome and all gene trees are concordant, while in scenario (**B)** the histories of the adaptive mutations differ from that of the rest of the genome, with gene trees that do not match the species tree.

4.3. Material and Methods

4.3.1. Sampling, sequencing and genome sizing

A low-coverage whole-genome sequencing approach (genome skimming) was used to reconstruct the genome history of *Alloteropsis*. This approach has become increasingly attractive for inferring population parameters (e.g. Buerkle and Gompert 2013; Fumagalli et al. 2013) and for studying complex traits (Li et al. 2011). It also allows de novo assembly of high copy number regions of the genome, such as organelle genomes (Besnard et al. 2014; Dodsworth 2015), and can be applied to samples of limited quality and quantity, such as herbarium or museum collections (Besnard et al. 2014). Genome-skimming data for eleven *A. semialata* individuals, and one of each of the congeneric *A. cimicina* and *A. angusta*, were retrieved from a previous study that used them to assemble chloroplast genomes (Table 4.S1; Lundgren et al. 2015, 2016). The photosynthetic type of these samples has been determined previously, and they

encompass non-C_4 individuals with and without a weak C_4 cycle, as well as multiple C_4 accessions (Table 4.S1; Lundgren et al. 2015, 2016). An additional eight *Alloteropsis* accessions were sampled here to increase the resolution of genome biogeography for the group (Table 4.S1). These include one accession from each of the congeneric species *A. paniculata* and *A. angusta*, and six additional *A. semialata* individuals. These samples were selected to increase the plastid and photosynthetic diversity, with a special focus on the Zambezian biogeographic region (spanning Tanzania, Zambia and the Democratic Republic of Congo – DRC; Linder et al. 2012; Table 4.S1), where the majority of the chloroplast and photosynthetic diversities are found (Lundgren et al. 2015, 2016). Three of the newly sequenced *A. semialata* accessions ('DRC3', 'TAN3' and 'KEN1') were previously characterized with stable carbon isotopes (Lundgren et al. 2015), which can distinguish plants grown using mainly C_4 photosynthesis from those that acquired a significant portion of their carbon via the ancestral C_3 cycle, whether or not it is complemented by a weak C_4 cycle (Smith and Brown 1973; Cerling et al. 1997). One of these three accessions ('TAN3') is isotopically intermediate, indicating that a strong C_4 cycle occurs, but that some atmospheric carbon is still fixed directly by the C_3 cycle (Peisker 1986; Monson et al. 1988). For four of the new samples, carbon isotopes were measured on a leaf fragment as previously described (Lundgren et al. 2015), which revealed that all of them had carbon isotope values within the C_4 range (Table 4.S1).

DNA was extracted, quality checked and sequenced as described in Lundgren et al. (2015), except that the DNA of these accessions was not sonicated prior to the library preparation due to the high degree of DNA degradation in these herbarium specimens. Each sample was individually barcoded and pooled with 23 other samples (from the same or unrelated projects) before paired-end sequencing (100–150 bp) on one Illumina lane (HiSeq-2500 or HiSeq-3000) at the Genopole platform of Toulouse or at the Genoscope platform of Evry (only *A. paniculata*; Table 4.S1). The final data set consisted of sequence data for a total of 21 individuals, sequenced in six different runs (Table 4.S1).

The genome size was estimated for accessions for which live material was available by flow cytometry following the one-step protocol of Doležel et al. (2007) with minor modifications as described in Clark et al. (2016). We selected *Oryza sativa* IR36 (2C = 1 pg; Bennett and Smith 1991) and the Ebihara buffer (Ebihara et al. 2005) as the most appropriate internal standard and nuclei isolation buffer for all but one

accessions (Table 4.S1). For the 'RSA3' accession, whose C-value was estimated to be about three time larger than other accessions, we used the Pisum sativum 'Ctirad' standard (2C = 9.09 pg; Doležel et al. 1992) and the GPB buffer (Loureiro et al. 2007), supplemented with 3% of PVP.

4.3.2. Assembly and analyses of chloroplast genomes

Complete chloroplast genomes were de novo assembled for the newly sequenced individuals using the genome walking method described in Lundgren et al. (2015). The newly generated chloroplast genomes were manually aligned with those already available, and a time-calibrated phylogenetic tree was inferred with beast v. 1.5.4 (Drummond and Rambaut 2007), as described in Lundgren et al. (2015). Monophyly of the outgroup (*A. cimicina* + *A. paniculata*) and the ingroup (*A. angusta* + *A. semialata*) was enforced to root the phylogeny, which is consistent with all previous analyses (Ibrahim et al. 2009; Christin et al. 2012; GPWGII 2012; Lundgren et al. 2015). The root of the tree was fixed to 11 Ma (as found by Lundgren et al. 2015), which was achieved with a normal distribution of mean of 11 and standard deviation of 0.0001. Two different analyses were run for 20 000 000 generations, sampling a tree every 1000 generations. After checking the convergence of the runs in tracer v. 1.5.0 (Drummond and Rambaut 2007), the burn-in period was set to 2 000 000 generations, and the maximum credibility tree was identified from the trees sampled after the burn-in period in both analyses, mapping median ages on nodes.

4.3.3. Genotyping across the nuclear genome

A reference genome for *Alloteropsis* is currently lacking. However, the grass *Setaria italica* (common name: Foxtail millet) belongs to the same tribe as *Alloteropsis* (Paniceae) and has a well-assembled reference genome (jgiv2.0.27; Bennetzen et al. 2012). *Setaria* and *Alloteropsis* diverged approximately 20 Ma (Christin et al. 2012), a time that is sufficient for a complete turnover of noncoding sequences (Ammiraju et al. 2008). However, reads corresponding to coding regions across the genome can still be reliably mapped (see Results).

Raw sequencing reads were quality filtered using the NGSQC toolkit v. 2.3.3 (Patel and Jain 2012). Reads with more than 20% of the bases having a quality score below Q20 and reads with ambiguous bases were removed. Furthermore, low-quality

bases (<Q20) were trimmed from the 3' end of the remaining reads. The filtered reads were mapped to the *Setaria* reference genome, using bowtie2 v. 2.2.3 (Langmead et al. 2009). Raw alignment files were cleaned using samtools v.1.2 (Li et al. 2009) and picard tools v.1.92 (http://picard.sourceforge.net/). PCR duplicates were removed, and only uniquely aligned reads in proper pairs were kept. This will remove most of the reads mapped to repetitive sequences, such as transposable elements, while retaining reads mapping to sequences that have been duplicated after the split of *Alloteropsis* and *Setaria*. The cleaned alignments were used to call single nucleotide polymorphic variants (SNPs) with samtools v. 0.1.19 using the mpileup function followed by the vcfutil.pl script with default setting supplied with the program. The South African C_4 individual 'RSA3' was excluded during SNP calling to avoid any bias that might result from the presence of more than two alleles in this polyploid (see Lundgren et al. 2015 and Table 4.S1). Genotypes of each accession, including 'RSA3', at all called SNP positions were extracted from the alignments using the mpileup function in samtools v.0.1.19, supplying the program with the positions of the called SNPs, and in-house developed scripts for further processing. The low-coverage data caused genotype probabilities to be low, which precluded effective filtering based on these probabilities. Therefore, fixed genotype calls were used. To evaluate the proportion of SNPs corresponding to exon sequences, annotations were extracted for the 25 727 coding regions of the *Setaria* genome with homologs in maize and rice genomes (from now on referred to as SZR homologs). The positions of the raw SNPs were intersected with the SZR homolog annotations in bedtools v.2.19.1 using default settings (Quinlan and Hall 2010).

SNPs with coverage above 2.5 times the genomewide coverage (Table 4.S2) were converted to unknown genotype calls. Furthermore, genotypes with more than two allele calls were also converted to missing data, and finally, positions with more than 50% missing data/unknown genotypes were discarded. The remaining 170 629 positions were used to infer a phylogenetic tree, using PhyML (Guindon et al. 2010) and a GTR substitution model (the best fit model as determined by hierarchical likelihood ratio tests), after coding heterozygous sites with IUPAC codes. Support was evaluated with 100-bootstrap pseudoreplicates. The low-coverage data likely cause some alleles to be missed, leading to an overestimate of homozygosity. However, no bias is expected in the missing allele, so that the low coverage is unlikely to lead to spurious groupings.

To test for a bias due to uneven coverage across samples (Table 4.S2), we repeated the phylogenetic analysis on a resampled alignment, where all samples have the same number of bases mapped to the *Setaria* genome. Reads were randomly sampled without replacement from the filtered alignment files until the number of bases across the sampled reads equalled that of the sample with the lowest coverage. These reanalyses were first conducted with all samples, which resulted in a low number of positions constrained to the samples with the lowest coverage. While analyses on the resampled data set were consistent with the whole-data set analyses, the limited number of characters resulted in reduced support. We consequently repeated the resampling allowing for the full alignment of the two *A. semialata* samples with the lowest coverage and alignment success ('AUS1' and 'RSA2') to be retained at a slightly lower coverage than the rest of the samples. SNPs were called as outlined above, which allowed for the retention of 22 821 SNPs.

4.3.4. Genetic structure and test for secondary gene flow

Preliminary cluster analyses with a focus on *A. semialata* showed that a more stringent filtering of the SNPs improved convergence of the analyses. Only positions with <10% missing data (2607 SNPs) within this species were consequently kept for analyses of its population structure, using the structure software (Pritchard et al. 2000). Ten independent analyses were run for each number of population components (K) from one to ten, under the admixture model. The adequate run length and burn-in periods were determined through preliminary analyses, which indicated that a burn-in period of 300 000 generations followed by 200 000 iterations provided stable estimates for all K values. The optimal K values were determined using the method of Evanno et al. (2005), as implemented in structureharvester (Earl and vonHoldt 2012). The results of the ten runs for each K were summarized using clumpp v. 1.1.2 (Jakobsson and Rosenberg 2007) and graphically displayed using distruct v. 1.1 (Rosenberg 2004). These analyses were repeated without the polyploid individual 'RSA3', which led to the same cluster assignments, showing that differences in ploidy levels do not affect the conclusions. Finally, the cluster analyses were repeated on alignments based on the reads subsampled to similar coverage in all samples, allowing for 25% missing data per site (retention of 681 SNPs).

Different relationships among fractions of the nuclear genome can result from reticulated evolution or incomplete lineage sorting (Green et al. 2010; Durand et al. 2011). To distinguish these two possibilities, the ABBA–BABA method, which relies on the D statistic to test for asymmetry in the frequencies of incongruent phylogenetic groupings (Green et al. 2010; Durand et al. 2011), was used to test for secondary gene flow on a genomewide level in cases suggested by phylogenetic and clustering analyses (see Results). The low coverage likely leads to an overestimate of homozygous sites, but no bias is expected towards ABBA or BABA sites, leaving estimations of distorted gene flow unaffected. For each test, a four-taxon phylogeny was selected, consisting of an outgroup and three tips among which secondary gene flow is suspected. Reads mapping to the 170 629 SNPs were recovered from the filtered alignment files using bedtools v.2.19.1 by intersecting the alignment files with positional information of the SNPs using default settings. The recovered reads were evaluated using the -doAbbababa option in the angsd program version 0.911 (Korneliussen et al. 2014). A jackknifed estimate of the D statistic and the corresponding Z-value were obtained by the jackknif.R script supplied with the angsd program.

4.3.5. Assembly and analyses of selected genes

Two different groups of closely related genes were selected for detailed analyses. The genes selected were two C_4-related protein-coding genes, phosphoenolpyruvate carboxylase (*ppc* genes) and phosphoenolpyruvate carboxykinase (*pck* genes), that include some copies acquired by *Alloteropsis* from distantly related species via lateral gene transfer, while other copies were vertically inherited following the species tree (Christin et al. 2012). Previous conclusions regarding the distribution of these genes among accessions of *Alloteropsis* were based on PCR and Sanger sequencing, which can be biased due to the possibility of primer binding mismatches. The presence/absence of the laterally acquired *ppc* and *pck* genes and their vertically inherited homologs across the accessions were therefore re-evaluated here using the genome-skimming data, as well as new PCR and Sanger sequencing with primer verified against the new genomic data. Using molecular dating, the divergence times of the laterally acquired genes were compared to those of vertically inherited homologs belonging to the same set of accessions.

Reads were first mapped on gene segments of the *ppc* and *pck* genes from different accessions of *Alloteropsis* (grass co-orthologs *ppc-1P3* and *pck-1P1*; Christin et al. 2012, 2015). These segments have been previously sequenced and analysed in a number of other C_3 and C_4 grasses (Christin et al. 2012). The availability of this rich reference data set allows mapping to closely related accessions of *Alloteropsis*, which improves the alignment success compared to the whole-genome approach described above, and increases the confidence in the assignment. The gene segments cover exons 8–10 for *ppc* and exons 3–10 for *pck*, including introns, and represent approximately 46% (1492 bp) and 63% (1487 bp), respectively, of the full-length coding sequences. In-house Perl scripts were used to unambiguously assign reads to genes of these data sets, following the phylogenetic annotation method of Christin et al. (2015). In summary, this approach consists of: (i) building a reference data set of sequences with known identity for closely related gene lineages, (ii) using blast searches to identify all sequences homologous to any of these reference sequences in the query data set (the filtered reads in this case), (iii) independently aligning each homologous sequence to the reference data set and inferring a phylogenetic tree and (iv) establishing the identity of each of the query sequences based on the phylogenetic group in which it is nested. Assignment of reads to the gene lineages was verified by visual inspection of the phylogenetic trees and the alignments. Subsequently, all reads assigned to each of the vertically inherited or laterally acquired gene lineages were retrieved, and aligned to PCR-isolated sequences (see Results) using geneious v. 6.8 (Kearse et al. 2012). The reads were then assembled into gene models, comprising introns and exons, for the studied segments. Multiple gene models were assembled for a single individual when the existence of distinct alleles was supported by at least two different polymorphic sites, each with at least two independent reads. Paired-end reads were then merged into contigs if they shared the polymorphisms. Reads that did not overlap the polymorphic sites were merged with all alleles, replacing additional polymorphisms with IUPAC ambiguity codes.

To check whether partial pseudogenes that do not include the studied segments exist in some genomes, the presence of laterally acquired *ppc* genes was also tested using only coding sequences corresponding to exons 1–7, which were retrieved from a transcriptome study of *A. semialata* (Christin et al. 2012). This transcriptome was generated for a South African C_4 polyploid with two laterally acquired *ppc* genes, but the vertically inherited versions of *ppc* and *pck* were not available in this transcriptome,

preventing phylogenetic analyses. Blastn searches were used to identify reads mapping to one of the two laterally acquired *ppc* genes on at least 50 bp with at least 99% of identity. Finally, the presence/absence of the different *pck* and *ppc* copies was further confirmed via PCR and Sanger sequencing using primers specific to the different gene copies (Table 4.S3; Christin et al. 2012). PCR, purification and sequencing were conducted as described in Lundgren et al. (2015), except for changes of the annealing temperature and/or extension time (Table 4.S3). These PCR were conducted only on samples for which good quality DNA was available. Indeed, DNA isolated from herbarium samples is highly degraded, precluding reliable PCR screening.

To verify the gene models assembled from genome skimming for *ppc* and *pck*, the PCR amplified and Sanger sequenced fragments of the vertically inherited and laterally acquired genes were added to the genes assembled from short-read data. The data sets were aligned using muscle v3.8.31 (Edgar 2004) with default parameters, and the alignments were manually refined. Maximum-likelihood phylogenetic trees were inferred using PhyML, under a GTR+G model, and with 100-bootstrap pseudoreplicates. Molecular dating was performed on the same alignments using beast as described above for chloroplast markers, but with a coalescent prior. The Andropogoneae/Paspaleae group (represented by *Sorghum*, *Paspalum* and one of the laterally acquired *ppc*) was selected as the outgroup, and the root of the tree was calibrated with a normal distribution with a mean of 31 Ma, and a standard deviation of 0.0001, as previously estimated for this node (Christin et al. 2014).

4.4. Results

4.4.1. Read alignment and SNP calling

The number of filtered paired-end reads varied across samples, for a genomewide coverage ranging from 0.70 to 4.52 (Table 4.S2). The proportion of filtered paired-end reads that aligned to the *Setaria* genome varied between 4.04% and 10.23%, and between 1.22% and 2.94% aligned to the coding regions (Table 4.S2). While the mapping was performed across the whole genome of *Setaria* (excluding the organelle genomes), divergence of noncoding sequences means that high mapping success is expected to be concentrated mostly onto coding sequences. About 9% of the *Setaria* genome corresponds to exons (Bennetzen et al. 2012). Assuming that the total length of exons is similar in the two species, the larger genome of *Alloteropsis* means that this

proportion should be about 4.5%, so that approximately half of the reads corresponding to nuclear exons were mapped. The rest of the reads that belong to exons probably correspond to gene sections that are too divergent between the two species to successfully map.

Only uniquely aligned reads were used to call SNPs, which inherently excludes common repetitive regions such as transposons. However, 1111 raw SNPs had a higher than expected coverage (>5×) across at least 50% of the samples. The positions of 91% (1007) of these high-coverage SNPs fell outside of the SZR homolog regions, and the rest were concentrated to only 14 SZR homologs. We therefore hypothesize that these high-coverage SNPs stem from genetic regions (mostly noncoding) that have been duplicated after the split between *Alloteropsis* and *Setaria* and they were subsequently removed from the analyses.

A total of 170 629 SNPs with <50% missing data across the 21 accessions were finally selected for downstream analyses. These sites are spread across all chromosomes (Fig. 4.S1) and 96% of them fall within one of 9948 SZR homologs. The 2607 SNPs used for the cluster analysis (<90% missing data across the *Alloteropsis semialata* samples) were equally well spread across the genome (Fig. 4.S1) and 97% fall within one of 848 SZR homologs. Most of the variation in missing data across samples (Table 4.S2) is likely explained by differences in coverage, although the presence/absence of genes within each accession might also influence the individual mapping success.

Overall, our analyses show that our pipeline, despite a low overall coverage and low alignment success due to the large divergence time between *Alloteropsis* and *Setaria*, captures variation in almost 10 000 genes spread across the genomes of grasses.

4.4.2. Phylogenetic trees

The plastid phylogeny identified two C_4 individuals from DRC with haplotypes that form a new C_4 plastid lineage based on divergence times (i.e. lineage G, sister to lineage F; Figs 4.2 and 4.S2). Relationships based on markers sampled across the nuclear genome confirm the monophyly of *A. semialata* and its sister-group relationship to *Alloteropsis angusta*, but present multiple incongruences with the chloroplast tree within *A. semialata* (Figs 4.2 and 4.S3). In this genomewide tree, the first divergence leads to a group composed of the non-C_4 accessions of *A. semialata* from South Africa without any known C_4 cycle (Clade I; Figs 4.2 and 4.S3), and the second divergence

leads to a group comprising the non-C_4 accessions from the Zambezian region that use a weak C_4 cycle (Clade II; Figs 4.2 and 4.S3; C_3-C_4 intermediates sensu Lundgren et al. 2016). The isotopically intermediate accession 'TAN3' is then sister to all C_4 accessions (Figs 4.2 and 4.S3). The two C_4 accessions bearing the plastid lineage G form a paraphyletic clade, while the other C_4 accessions from the Zambezian region ('TAN4', 'DRC3' and 'DRC4') are grouped in a strongly supported clade (Clade III; Figs 4.2 and 4.S3). The South African polyploid individual 'RSA3' is sister to the C_4 individuals sampled outside of the Zambezian region, and the rest of the C_4 accessions form the strongly supported clade IV, with two subclades corresponding to Africa plus Madagascar and Asia plus Australia (Figs 4.2 and 4.S3). The nuclear phylogeny based on the resampled data set is mostly identical to the one based on the whole data set (Figs 4.S3 and 4.S4).

4.4.3. Genetic structure and secondary gene flow within Alloteropsis semialata

Based on the whole-genome clustering analysis, four clusters explain the data set best, and adding groups does not improve the likelihood (Fig. 4.3B). However, the method of Evanno et al. (2005) indicates that the maximum fit improvement is at two clusters, with four clusters representing the second maximum fit improvement (Fig. 4.3C). With four clusters, the main clades from the genome wide phylogeny are recovered (Figs 4.2 and 4.3A). This genetic structure matches the photosynthetic types rather than the geographic origin, with the non-C_4 clades I and II and the C_4 clades III and IV each forming distinct homogenous groups (Fig. 4.3A). The three Zambezian individuals that formed a paraphyletic clade in the nuclear phylogeny ('TAN3', 'DRC1' and 'DRC2') are partially assigned to two Zambezian groups, the non-C_4 clade II and the C_4 clade III (Figs 4.2 and 4.3A). Finally, the polyploid individual from South Africa, 'RSA3', is partially assigned to the two C_4 clades III and IV (Fig. 4.3A). The cluster results based on the resampled data set are less stable due to a lower number of sites, but the assignments are similar (Figs 4.3 and 4.S5).

Heterozygosity was estimated for each sample based on the 22821 SNPs from the resampled data set with similar coverage across samples. While these estimates are based only on sites variable within *Alloteropsis* and should consequently not be interpreted as genomewide heterozygosity, it is possible to compare the estimates

between samples. The individuals assigned to multiple clusters had the highest percentage of heterozygous SNPs (Fig. S6), which confirms their genetic diversity.

Together, our intraspecific genetic analyses reveal the existence of distinct gene pools despite overlapping distributions (Figs 4.3A and 4.4), but also suggest that genetic exchanges have happened among groups. The incongruences between the phylogenetic structures of the chloroplast and nuclear genomes, together with the assignment of some individuals to multiple genetic clusters, suggest that the three Zambezian individuals 'TAN3', 'DRC2' and 'DRC1' have ancestors from distinct genetic groups, in this case the nuclear clades II and III. ABBA–BABA tests were therefore conducted to test this hypothesis, using *A. angusta* (individual Ang2) as the outgroup. The individual 'TAN4' was selected as the representative of clade III because it is geographically more distant and distinct on all genetic markers (Figs 4.4, 4.S2 and 4.S3). Significant indications ($P < 0.05$ after correction for multiple testing) of gene flow from the non-C_4 clade II ('TAN2' and 'TAN1') into the populations assigned to multiple clusters ('TAN3', 'DRC2' and 'DRC1') were found (Table 4.1). In contrast, there is no evidence of a significant secondary contribution of clade II into individuals of clade III ('DRC3' or 'DRC4'; Table 4.1). However, in one case, a slight excess of BABA sites was detected, which was not significant after correction for multiple testing (Table 4.1). This would suggest some genetic contribution from one non-C_4 population of clade II ('TAN1') into the C_4 population represented by 'TAN4' (Table 4.1).

Within clade IV, the C_4 individual from Madagascar ('MAD1') was grouped with Asian C_4 accessions on plastid genomes but grouped with the African C_4 accessions based on markers from across the nuclear genome (Figs 4.S2 and 4.S3). An ABBA–BABA test was conducted to test the hypothesis of secondary gene flow after the split of the African and Asian C_4 accessions. The accession 'TAN4' was used as the outgroup, being sister to all accessions from clade IV. The Taiwan accession ('TPE1') was selected as the Asian sample, while the Burkinabe accession ('BUR1') represented Africa. Overall, more ABBA than BABA sites were detected (Table 4.1), indicating that the Asian accession was closer to the accession from Madagascar ('MAD1') than to the accession from mainland Africa, but the D statistic for this test was not significant after correcting for multiple testing (Table 4.1). Plastid markers, which represent seed dispersal, group the Madagascan accessions with Asian individuals. Therefore, a possible scenario involves an initial seed dispersal from Africa to Madagascar and then from Madagascar to Asia, with subsequent pollen flow between Africa and Madagascar.

4.4.4. Assembly and analyses of selected genes

The presence/absence of *ppc* and *pck* genes was established by mapping reads directly onto reference sequences from *Alloteropsis*. The distribution of the genes was also confirmed by PCR followed by Sanger sequencing (Fig. 4.S7).

Together, the results confirmed previous phylogenetic analyses (Christin et al. 2012), but with a significant increase of the sample size. Reads assigned to the *pck* gene copy laterally acquired from members of the *Cenchrus* genus (*pck-1P1_LGT:C*) were detected in the two *A. angusta* accessions and all *A. semialata* accessions, except the two non-C_4 accessions from South Africa (Table 4.2; Figs 4.S7 and 4.S8). The sequences assembled for the laterally acquired *pck* gene were highly similar between the different accessions, leading to a poorly resolved phylogeny (Fig. 4.S8). By contrast, the sequences assembled for the vertically inherited *pck* gene were variable among accessions, and the nuclear clades I and II were recovered in their phylogeny, while clades III and IV were not differentiated (Fig. 4.S8). Interestingly, one accession with mixed genetic backgrounds ('DRC1') has two divergent alleles, one of which groups with clade II and the other with clade III/IV (Fig. 4.S8). Dating analyses indicate that the divergence of *A. angusta* and *A. semialata* is more recent for the laterally acquired *pck* than for the vertically inherited copy (Fig. 4.S9). However, the divergence of C_4 accessions of *A. semialata* is estimated at a similar time based on the vertically inherited and laterally acquired *pck* (Fig. 4.5).

The vertically inherited *ppc* was recovered from all samples, and the assembled gene models were variable enough to partially resolve the phylogeny, with well-supported clades corresponding to the different species (Fig. 4.S10). Although support was limited within *A. semialata*, the non-C_4 clades I and II (including sequences from individuals assigned to multiple clades) were sister to a clade composed of the C_4 accessions from clade IV nested within those of clade III (Fig. 4.S10). The divergence of vertically inherited *ppc* from C_4 accessions (excluding those partially assigned to clusters other than III and IV) matches the divergence of the vertically inherited *pck* for the same accessions (Fig. 4.5).

The *ppc* gene laterally acquired from Andropogoneae (*ppc-1P3_LGT:A*) was only detected in the Australian C_4 accession ('AUS1'; Table 4.2, Figs 4.S7 and 4.S10). An almost complete sequence for the studied segment was assembled, which was identical to those isolated by PCR.

The *ppc* gene laterally acquired from the *Setaria palmifolia* complex (*ppc-1P3_LGT:C*) was detected in the C_4 accessions from South Africa ('RSA3') and the DRC (Table 4.2, Figs 4.S7 and 4.S10). Although no reads matching exons 8–10 of *ppc-1P3_LGT:C* were detected in the accession 'TAN3', a total of seven reads from this individual matched exons 1–7. This suggests that the gene is truncated and probably exists as a pseudogene in this individual. The *ppc-1P3_LGT:C* sequences were largely conserved, although distinct alleles were assembled in one of the accessions with mixed genetic background ('DRC2'; Fig. 4.S10). The divergence of *ppc-1P3_LGT:C* genes belonging to different C_4 accessions was more recent than for the vertically inherited *ppc* and *pck* of the same accessions (Fig. 4.5).

The *ppc* gene acquired from Melinidinae (*ppc-1P3_LGT:M*) was identified in nine C_4 accessions of *A. semialata*, the isotopically intermediate *A. semialata*, and the two congeners *A. cimicina* and *A. paniculata* (Table 4.2, Figs 4.S7 and 4.S10). Highly divergent alleles of the *ppc-1P3_LGT:M* gene were inferred for *A. cimicina* and *A. paniculata* (Fig. 4.S10). However, the sequences of *ppc-1P3_LGT:M* from *A. semialata* were very similar to each other, and nested within the alleles from *A. cimicina*/*paniculata* (Fig. 4.S10). The split of *A. semialata* and *A. cimicina* is more recent for *ppc-1P3_LGT:M* than for the vertically inherited *ppc* and *pck* (Fig. 4.S9). In addition, the divergence of C_4 accessions of *A. semialata* based on this *ppc-1P3_LGT:M* gene occurred more recently than the divergence based on the vertically inherited *ppc* and *pck* (Fig. 4.5).

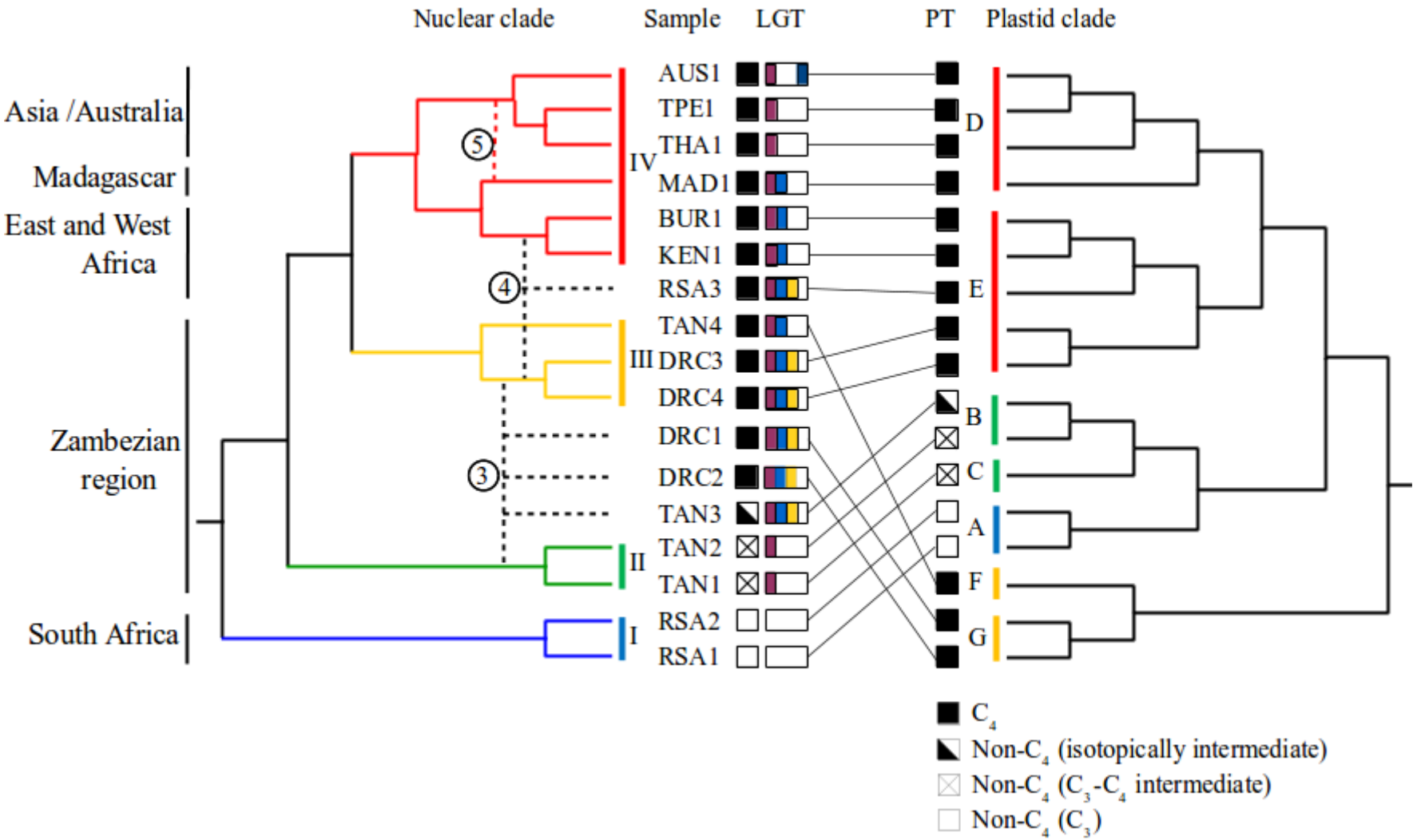


Fig. 4.2. Comparison of nuclear (on the left) and plastid (on the right) phylogenetic topologies (without branch lengths). The putative origin of individuals with mixed genetic background was added using dashed lines. Branches of the nuclear tree are coloured according to clustering analyses (Fig. 4.3). Photosynthetic types (PT) and presence of laterally acquired genes (LGT) are indicated by symbols at the tips; purple bar = presence of *pck-1P1_LGT:C*, blue bar = *ppc-1P1_LGT:M*, orange bar = *ppc-1P1_LGT:C*, dark blue bar = *ppc-1P1_LGT:A*. Geographic origin is indicated on the left. Secondary gene flow is numbered as in Fig. 4.6; (3) hybridization between non-C_4 and C_4 populations within the Zambezian region, (4) alloplyploid between C_4 populations in African ('RSA3'), (5) pollen-mediated gene flow from mainland Africa to Madagascar.

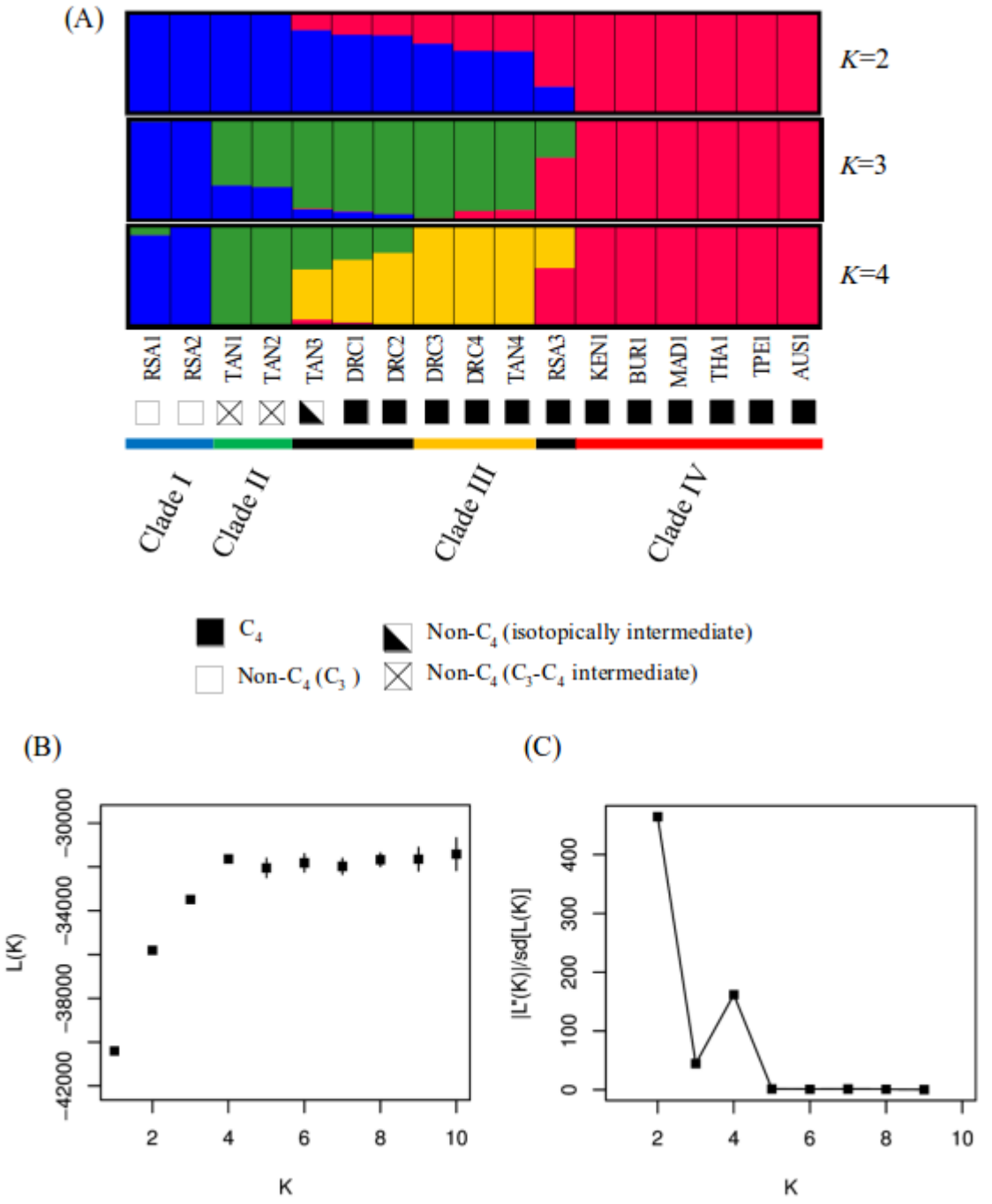


Fig. 4.3. Assignment of *Alloteropsis semialata* individuals to genetic clusters. (A) Assignment of each individual to the different clusters (*K* 2–4). The photosynthetic type is indicated by symbols next to the names, as in Fig. 4.2; (B) Mean likelihood (±SD) over 10 runs for each *K* value (1–10); (C) |L"(K)|/SD (fit improvement) as calculated according to Evanno et al. (2005).

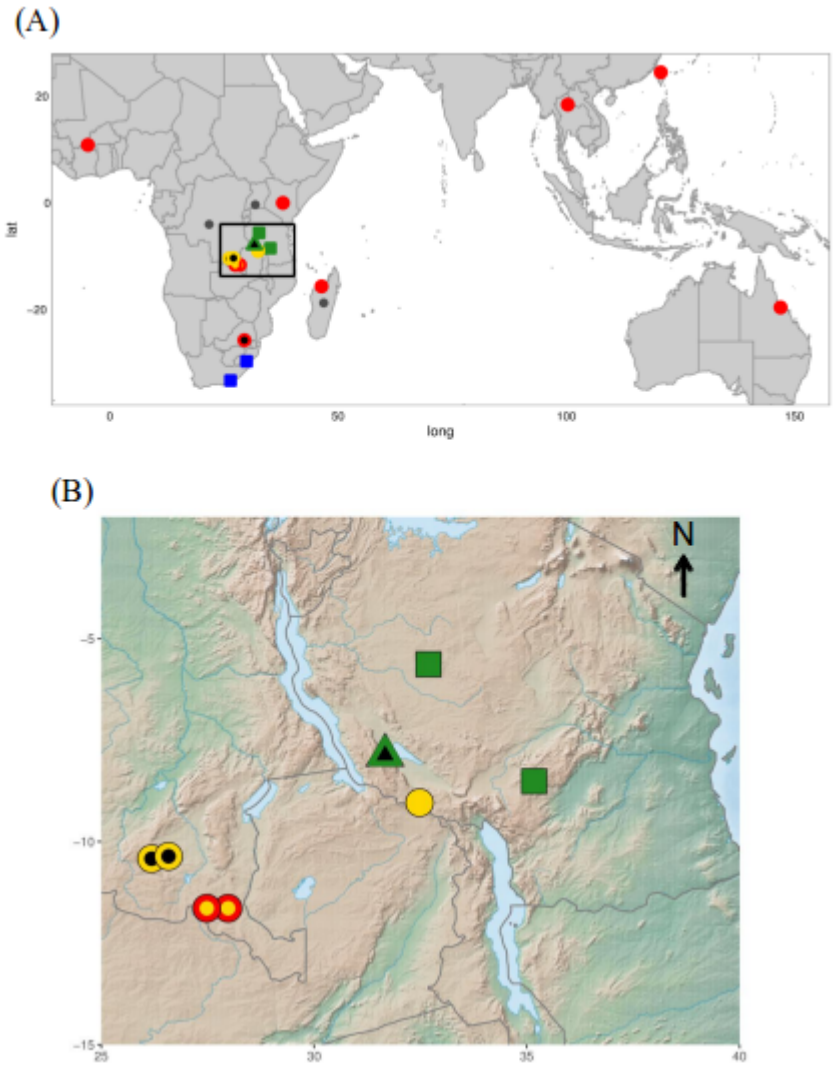


Fig. 4.4. Geographic distribution of *Alloteropsis semialata* genetic lineages. (A) World distribution, highlighting the Zambezian region with a rectangle and (B) details of the Zambezian region. For each point, the colour of the outline indicates the plastid lineage (blue = clade A; green = clade BC; yellow = clade FG; red = clade DE), while the colour of the background represents the nuclear lineage (blue = clade I; green = clade II; yellow = clade III; red = clade IV; black = mixed genetic background; grey = congeners). Finally, the shape of the point indicates the photosynthetic type, as determined by carbon isotopes (square = non-C_4; circle = C_4; triangle = isotopically intermediate).

Table 4.1. Results of ABBA-BABA tests.

Outgroup[1]	P3[1]	P2[1]	P1[1]	# ABBA sites	# BABA sites	D[2]	Z	*P*-value[3]	Conclusion
A. angusta	TAN2	TAN3	TAN4	2630	1805	0.186	8.279	<0.0001	TAN2 closer to TAN3 than to TAN4
A. angusta	TAN1	TAN3	TAN4	2630	1750	0.201	8.757	<0.0001	TAN1 closer to TAN3 than to TAN4
A. angusta	TAN2	DRC2	TAN4	2037	1546	0.137	5.939	<0.0001	TAN2 closer to DRC2 than to TAN4
A. angusta	TAN1	DRC2	TAN4	1960	1570	0.110	4.724	<0.0001	TAN1 closer to DRC2 than to TAN4
A. angusta	TAN2	DRC1	TAN4	2240	1752	0.122	5.463	<0.0001	TAN2 closer to DRC1 than to TAN4
A. angusta	TAN1	DRC1	TAN4	2194	1749	0.113	5.692	<0.0001	TAN1 closer to DRC1 than to TAN4
A. angusta	TAN2	DRC4	TAN4	1177	1164	0.006	0.223	0.824	TAN2 equally close to DRC4 and TAN4/correct phylogeny
A. angusta	TAN1	DRC4	TAN4	1075	1123	-0.022	-0.866	0.386	TAN1 equally close to DRC4 and TAN4/correct phylogeny
A. angusta	TAN2	DRC3	TAN4	1372	1451	-0.028	-1.080	0.280	TAN2 equally closer to DRC3 and TAN4/correct phylogeny
A. angusta	TAN1	DRC3	TAN4	1248	1431	-0.068	-2.885	0.004[4]	TAN1 might be closer to TAN4 than to DRC3
TAN4	TPE1	MAD1	BUR1	1314	1129	0.076	2.603	0.009[4]	TPE1 might be closer to MAD1 than to BUR1

[1] (Outgroup, (P3, (P2, P1))).
[2] D statistic: (ABBA-BABA)/(ABBA+BABA).
[3] *P*-value for Z score as calculated by jackknife for whether D differs significantly from zero.
[4] Nonsignificant after Bonferroni correction for multiple testing.

Table 4.2. Number of reads assigned to each of the laterally acquired *pck* and *ppc* genes.

Species	Accession	Phylogenetic group (plastid; nuclear)	*ppc-1P3 LGT:A* [1]	*ppc-1P3 LGT:M* [2]	*ppc-1P3 LGT:C* [3]	*pck-1P1 LGT:C* [3]
A. cimicina	Cim1	-	0	149[4]	0	0
A. paniculata	Pan1	-	0	37[4]	0	0
A. angusta	Ang2	-	0	0	0	78
	Ang1	-	0	0	0	49
A. semialata	RSA1	A; I	0	0	0	0
	RSA2	A; I	0	0	0	0
	TAN1	C; II	0	0	0	57
	TAN2	B; II	0	0	0	73
	TAN3	B; mixed	0	54[4]	0[5]	216[4]
	DRC1	G; mixed	0	57[4]	56	183[4]
	DRC2	G; mixed	0	29	50[4]	95[4]
	DRC3	E; III	0	6	25	135[4]
	DRC4	E; III	0	10	12	88[4]
	TAN4	F; III	0	76	0	83
	RSA3	E; IV	0	55[4]	63	113
	KEN1	E; IV	0	36	0	85
	BUR1	E; IV	0	26	0	130
	MAD1	D; IV	0	46	0	101
	THA1	D; IV	0	0	0	123
	TPE1	D; IV	0	0	0	118
	AUS1	D; IV	55	0	0	110

[1] Laterally acquired from Andropogoneae.
[2] Laterally acquired from Melinidinae.
[3] Laterally acquired from Cenchrinae.
[4] Assembly of more than one allele.
[5] Note that seven reads were retrieved for exons 1-7, which indicates that this gene is truncated in the genome of this accession.

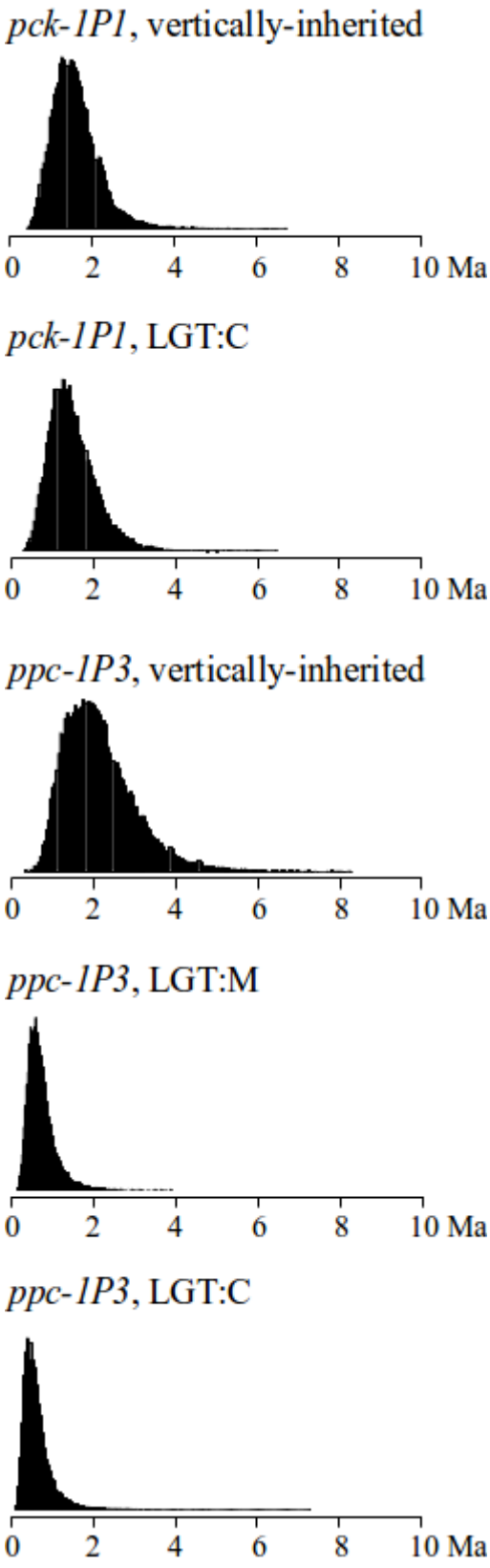


Fig. 4.5. Divergence times of C_4 accessions of *Alloteropsis semialata* based on vertically inherited and laterally acquired genes. For five *ppc* and *pck* genes, the posterior distribution of times to the last common ancestor of the C_4 *A. semialata* is shown, in million years (Ma).

4.5. Discussion

4.5.1. Divergence of photosynthetic types in isolation followed by secondary gene flow

Overall, our genomewide analyses reveal a strong genetic structure, which matches photosynthetic types better than geographic origins, although both play a role. All C_4 individuals form a monophyletic group based on genomewide markers, which is sister to a clade composed of non-C_4 accessions from the Zambezian region with a weak C_4 cycle (clade II; Figs 4.2 and 4.4), and together, these two groups are sister to the non-C_4 accessions from South Africa that lack a C_4 cycle (clade I; Figs 4.2 and 4.4). The C_4 clade contains two clearly distinct subgroups, one from the Zambezian region (clade III; Figs 4.2 and 4.4) and the other one encompassing all C_4 accessions sampled outside this region, from Western Africa to Australia (clade IV; Figs 4.2 and 4.4). The Zambezian region encompasses more genetic diversity than the rest of the species' range, including a total of five plastid lineages, four of which are endemic (clades B, C, F and G; Figs 4.2 and 4.S2). This finding further supports this region as the centre of origin for *Alloteropsis semialata* (Lundgren et al. 2015). Based on both plastid and nuclear genomes, the divergence of photosynthetic types likely also happened within this region (Fig. 4.6). Both C_4 and non-C_4 populations in the Zambezian region are associated with Miombo woodlands. Periodic cycles of contraction and expansion of these wooded savannas during recent geological times might have isolated populations of *A. semialata* in this geologically and topographically complex region (Cohen et al. 2007; Beuning et al. 2011). The ancestral photosynthetic state is likely non-C_4 and mutations altering the leaf anatomy and upregulation of enzymes already present in the non-C_4, ancestors likely led to the emergence of a constitutive C_4 cycle in some isolated populations (Mallmann et al. 2014; Bräutigam and Gowik 2016). One of the lineages descending from the initial C_4 pool, corresponding to clade IV, later left the Miombo of the Zambezian region and rapidly spread across Africa and all the way to Asia and Australia (Figs 4.4 and 4.6). This biogeographical history therefore points to the initial emergence of the C_4 physiology in *A. semialata* within the Zambezian region, with subsequent isolation of the C_4 descendants (Fig. 4.6).

The lack of association between chloroplast and nuclear groups (Figs 4.2, 4.S2 and 4.S3) in the Zambezian region suggests ancient, but recurrent, secondary gene flow

followed by homogenization of the local gene pools. In addition, the presence of three individuals with mixed nuclear backgrounds indicates relatively recent gene flow between previously isolated groups. The maximum expansion of the Miombo woodlands during interglacial periods, as presently occurs, would likely favour seed dispersal over a larger area, leading to secondary contacts (Vincens 1989; Cohen et al. 2007; Beuning et al. 2011), a process frequently reported in Europe (reviewed in, e.g. Hewitt 2000; Schmitt 2007). We propose that this expansion allowed genetic exchanges between previously isolated lineages, some of which had made the transition to a full C_4 physiology during the previous isolation. No evidence of gene flow between C_4 and non-C_4 individuals was found outside of the Zambezian region, and crosses might be prevented in South Africa, the other region where C_4 and non-C_4 individuals overlap, by differences in ploidy levels (Fig. 4.4; Liebenberg and Fossey 2001). However, our analyses suggest that allopolyploidy contributed to the mixing of nuclear groups III and IV in Southern Africa (Fig. 4.6). In addition, while the recent divergence decreases statistical confidence, we found suggestions for secondary gene flow between different subgroups of the C_4 clade IV in Madagascar (Fig. 4.6).

Repeated isolation followed by recurrent, but rare secondary gene flow has created a dynamic population structure whereby adaptive mutations, such as those for the C_4 trait, can appear and sweep to fixation in isolation and later come together through admixing in times of contact. While mutations for increasing the expression of the C_4-related genes and altering the leaf anatomy are unknown, genes for two key C_4 enzymes were laterally acquired by *A. semialata* (Christin et al. 2012). These lateral gene transfers likely took place in *A. semialata* plants that already used C_4 photosynthesis, and once acquired, these genes presumably replaced the function of the vertically inherited gene copies that were overexpressed but not biochemically optimized (Christin et al. 2012). The biogeographic history inferred here for the nuclear genome allows us to estimate the region where these lateral gene transfers likely occurred and track the subsequent spread of these genes among different gene pools.

4.5.2. Spread of C_4-adaptive mutations among gene pools

Our analyses detected the laterally acquired *pck* gene in all *Alloteropsis angusta* and *A. semialata* individuals apart from two non-C_4 *A. semialata* South African accessions of *A. semialata* from South Africa, confirming previous PCR-based approaches (Table 4.2;

Christin et al. 2012). The divergence time is younger between the laterally acquired *pck* genes from *A. angusta* and *A. semialata* than between the vertically inherited genes of the same species (Figs 4.5 and 4.S9). This suggests that the laterally acquired *pck* was passed between *A. angusta* and *A. semialata* through secondary gene flow.

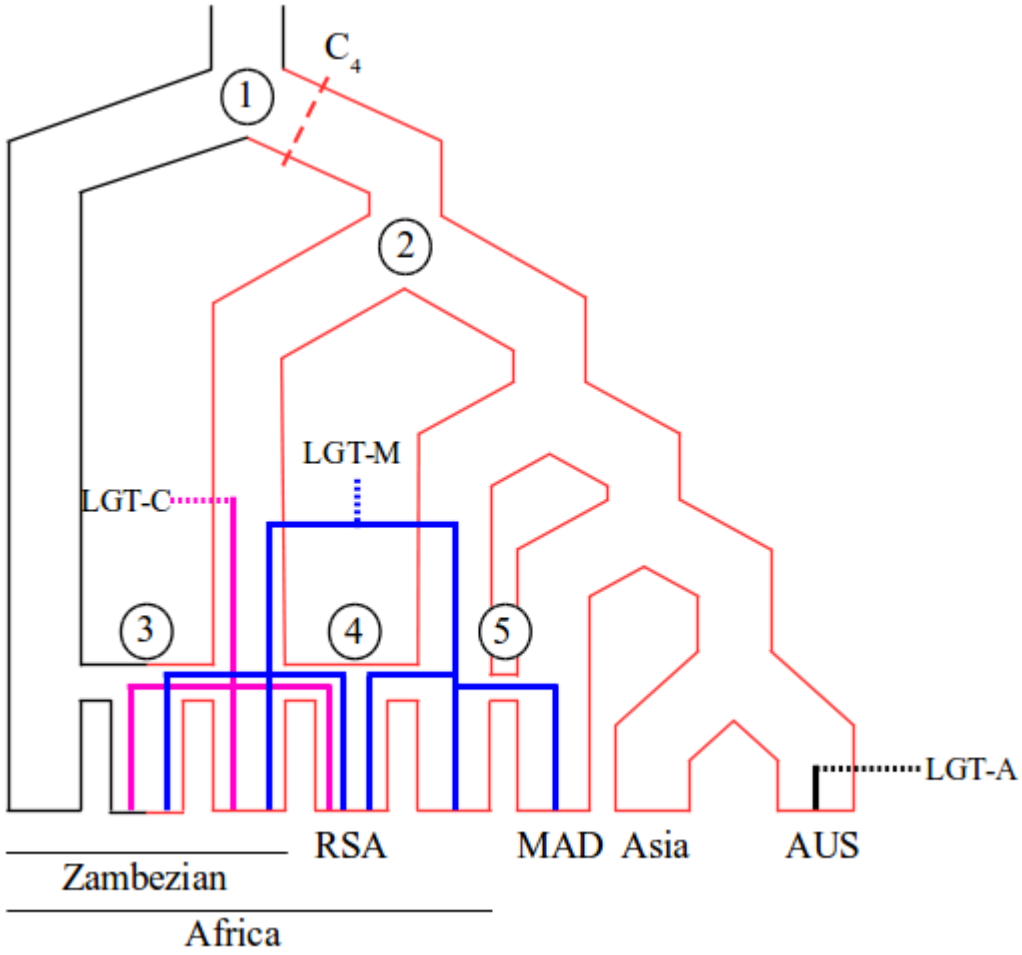


Fig. 4.6. Inferred history of divergence, secondary exchanges and spread of laterally acquired *ppc* genes in *A. semialata*. A summary phylogeny is shown for the C_4 and non-C_4 accessions of *A. semialata*, excluding the non-C_4 individuals from South Africa. The C_4 phenotype is represented with red outlines. (1) The divergence of photosynthetic types is inferred in the Zambezian region (dashed red line indicates C_4 emergence). (2) A C_4 lineage migrated outside of the Zambezian region. (3) Hybridization occurred between non-C_4 and C_4 populations within the Zambezian region. (4) The C_4 polyploids of South Africa (RSA) resulted from segmental allopolyploidy. (5) Pollen-mediated gene flow occurred from mainland Africa to Madagascar. The lateral acquisition of three ppc genes is indicated with dashed lines, and their subsequent spread is indicated with solid lines. Geographic regions are indicated at the bottom.

The accessions from Taiwan and Thailand do not possess any laterally acquired *ppc* genes, yet carbon isotopes unambiguously indicate that they carry out C_4 photosynthesis (Table 4.2; Lundgren et al. 2015). It is therefore likely that they overexpress their vertically inherited *ppc* and other genes required to generate a working C_4 cycle in the absence of repeated rounds of fixation of adaptive amino acids, as

observed in older C_4 lineages (Christin et al. 2007; Besnard et al. 2009; Huang et al. in press).

Out of the three different *ppc* genes acquired via lateral gene transfers from distant C_4 relatives (Table 4.2; Christin et al. 2012), only *ppc-1P3_LGT:A* is restricted to one of the accessions sampled here ('AUS1'). This gene was only found in Australia, and it is thus likely that it was recently acquired in this region (Fig. 4.6). The other two laterally acquired *ppc* genes are absent from some individuals, but spread across multiple populations of *A. semialata* that belong to different genomic clusters (Table 4.2). This pattern could result from the presence of the gene in the common ancestor and subsequent losses in some populations. However, this scenario is not supported by the lack of phylogenetic structure on the laterally acquired genes (Fig. 4.S10) and the comparison of divergence times, which indicate that the divergence of variants of both *ppc-1P3_LGT:M* and *ppc-1P3_LGT:C* found in C_4 accessions is more recent than the divergence of vertically inherited genes in the same accessions (Fig. 4.5).

The laterally acquired *ppc-1P3_LGT:M* gene was identified in the C_4 congeners *Alloteropsis cimicina* and *Alloteropsis paniculata*, as well as all C_4 accessions of *A. semialata* from Africa and Madagascar (whether from clade III or IV; Table 4.2). However, this gene was absent from the Asian/Australian C_4 accessions from clade IV and the African non-C_4 (clades I and II; Table 4.2). The divergence time between *ppc-1P3_LGT:M* genes belonging to *A. cimicina* and *A. semialata* is younger than the divergence times for the vertically inherited genes from the same species (Fig. 4.S9). In addition, the higher allelic diversity in *A. cimicina* compared to *A. semialata* suggests that the *ppc-1P3_LGT:M* gene was first acquired by *A. cimicina* and then transferred to *A. semialata*, potentially via hybridization. This gene has subsequently likely spread across distinct genetic groups of *A. semialata* in Africa and Madagascar via secondary pollen flow (Fig. 4.6). The fixation of the *ppc-1P3_LGT:M* gene within different populations would have been favoured by its improvement of the C_4 cycle, a function for which it was already optimized after millions of years spent in another C_4 lineage. Once this adaptive gene copy was acquired in a population, the vertically inherited *ppc* copy probably underwent pseudogenization as a result of relaxed selection. Indeed, the vertically inherited *ppc* genes bear frameshift mutations causing loss of function in two accessions with the laterally acquired *ppc-1P3_LGT:M* ('TAN4' and 'Cim1'), supporting the hypothesis that their function was taken over by the newly acquired gene, making them obsolete.

The last of the laterally acquired *ppc* genes, *ppc-1P3_LGT:C*, was found in the South African C_4 polyploid ('RSA3') as well as in four C_4 and one isotopically intermediate individuals from the Zambezian region, two from clade III and three with genetic contributions from clades II and III (Table 4.2). This gene was laterally acquired from a species of the *Setaria palmifolia* complex (Christin et al. 2012), which co-occurs with *A. semialata* in Zambezian Africa, where they grow metres apart, but not in South Africa (Clayton 1979). The transfer therefore likely occurred in the Zambezian region and later spread among the C_4 populations in this region through secondary gene flow (Fig. 4.6). Once acquired the *ppc-1P3_LGT:C* gene presumably took over the C_4 function, which might have been fulfilled by the previously acquired *ppc-1P3_LGT:M*. Indeed, *ppc-1P3_LGT:M* is still expressed in the transcriptome of the South African C_4 accession, but possesses internal stop codons that prevent proper translation (Christin et al. 2012). The newly acquired *ppc-1P3_LGT:C* likely spread to the C_4 populations from South Africa, through the putative segmental allopolyploidy event, providing a mechanism to propagate adaptive loci across genetic pools (Fig. 4.6). However, the Melinidinae *ppc-1P3_LGT:M* discussed above was spread among diploid individuals from clades III and IV, showing that adaptive loci can be transmitted despite limited gene flow, without the need for polyploidization.

The laterally acquired genes, which can easily be tracked using genome scans, show that the distinct genetic pools in *A. semialata* constitute reservoirs of genes for the adaptation of other populations within the same species complex. The history of these markers proves that genes for a complex trait can evolve independently in isolated populations and later be combined via natural selection following gene flow. When high-quality genome data accumulate for multiple accessions of *A. semialata*, such a scenario can be tested for vertically inherited genes, potentially explaining how novel adaptations can evolve in fragmented species complexes.

4.6. Conclusions

In this study, we analysed genomic data from multiple accessions of the grass *Alloteropsis semialata* using low-coverage whole-genome sequencing. Using a biogeographic framework for different parts of the genome, we demonstrate that multiple genetic pools exist, which are generally associated with different photosynthetic types. These pools originated more than 2 million years ago in the

Zambezian region and were kept relatively isolated, but with recurrent secondary gene flow, including between non-C_4 and C_4 individuals. These genetic exchanges contributed to the spread of adaptive loci, as illustrated by key C_4 genes acquired laterally in the Zambezian region and then rapidly passed to other African C_4 accessions. This process likely gradually optimized the initial C_4 pathway of some *A. semialata* populations through the assembly of different components. These genetic elements evolved in different parts of the species range, where limited gene flow might have facilitated local adaptation, but their subsequent combination likely improved the efficiency of the photosynthetic pathway of some accessions.

4.7. Acknowledgements

This work was funded by a Royal Society University Research Fellowship URF120119, a NERC grant NE/M00208X/1, an ERC grant ERC-2014-STG-638333 to PAC and a 'Ciência sem Fronteiras' CNPq scholarship 201873/2014-1 to MB. GB is supported by the 'Laboratoire d'Excellence (LABEX)' entitled TULIP (ANR-10-LABX-0041; ANR-11-IDEX-0002-02) and received support from the PhyloAlps project. The authors thank the Royal Botanic Gardens, Kew, the Botanic Garden Meise, and the National Museums of Kenya, Nairobi, which provided the samples used in this study. Olivier Bouchez from the Genopole in Toulouse helped with the Illumina sequencing.

4.8. Data accessibility

All raw reads are available in the short sequence archive under Accession no. SRP082653. In the NCBI nucleotide database, all newly assembled chloroplast genomes are available under Accession nos KX752083- KX752090, ppc genes under Accession nos KX788072- KX788087 and pck genes under Accession nos KX788088-KX788109.

4.9. Supporting Information

Figure 4.S1. Distribution of the 171,908 called SNPs along the *Setaria italica* genome.

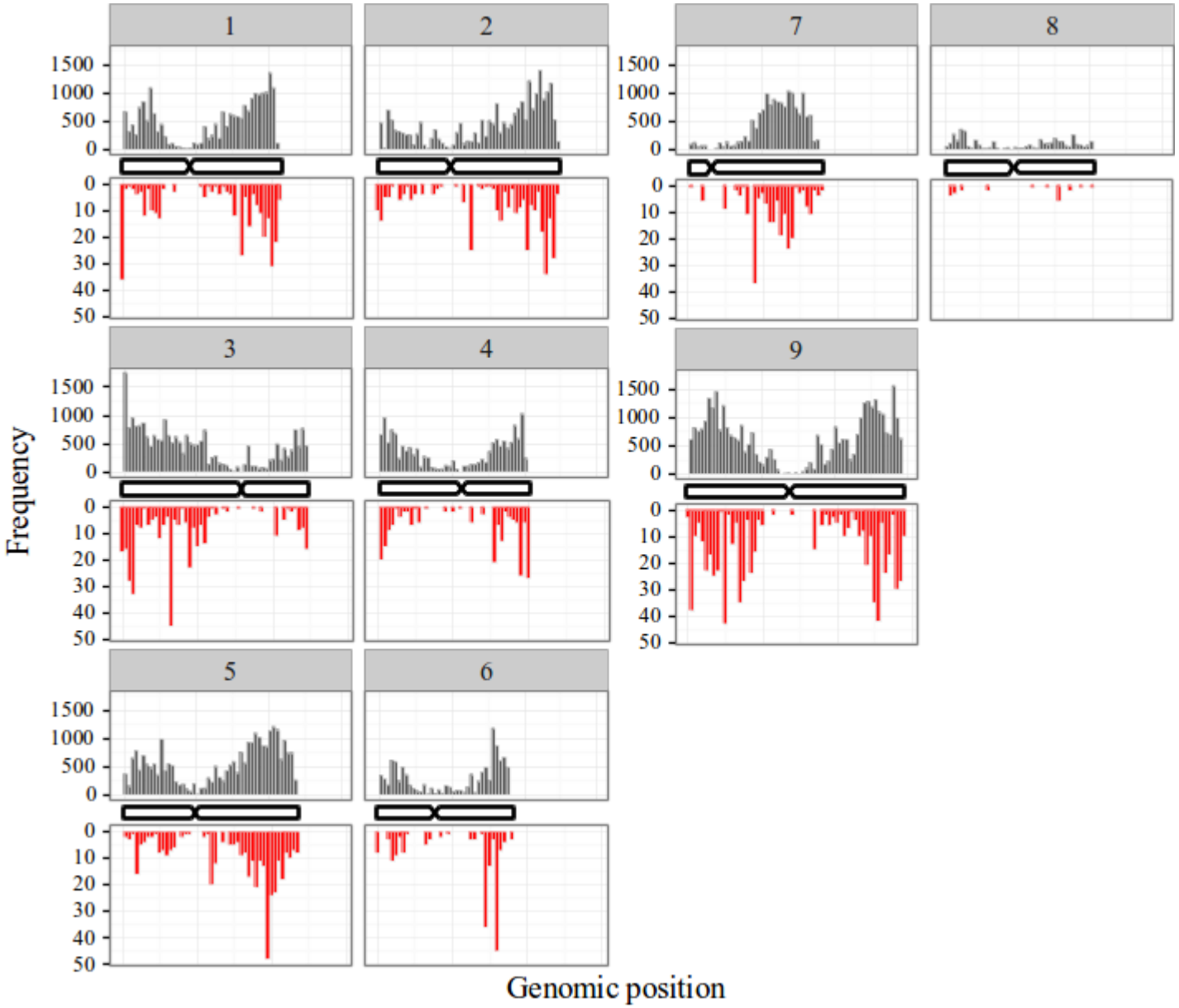

1
2
7
8
3
4
9
5
6
1500
1000
500
0
0
10
20
30
40
50
Frequency
Genomic position

Fig. 4.S2. Phylogenetic relationships based on complete chloroplast genomes from *Alloteropsis*.

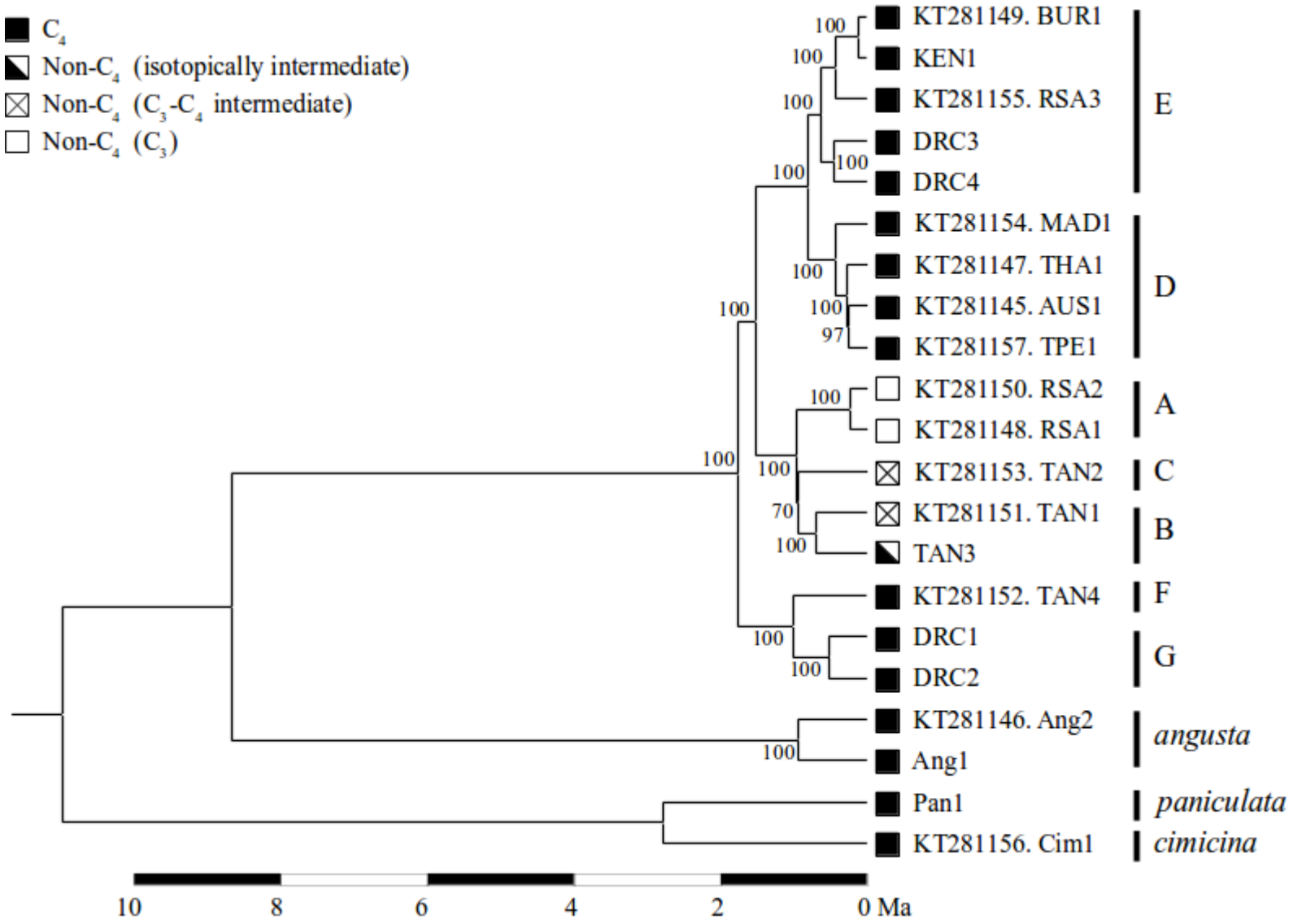

C4
Non-C4 (isotopically intermediate)
Non-C4 (C3-C4 intermediate)
Non-C4 (C3)
KT281149. BUR1
KEN1
KT281155. RSA3
DRC3
DRC4
KT281154. MAD1
KT281147. THA1
KT281145. AUS1
KT281157. TPE1
KT281150. RSA2
KT281148. RSA1
KT281153. TAN2
KT281151. TAN1
TAN3
KT281152. TAN4
DRC1
DRC2
KT281146. Ang2
Ang1
Pan1
KT281156. Cim1
E
D
A
C
B
F
G
angusta
paniculata
cimicina
100
100
100
100
100
100
100
97
100
100
70
100
100
100
100
10
8
6
4
2
0 Ma

Fig. 4.S3. Phylogenetic relationships based on whole genome sequencing of *Alloteropsis* accessions.

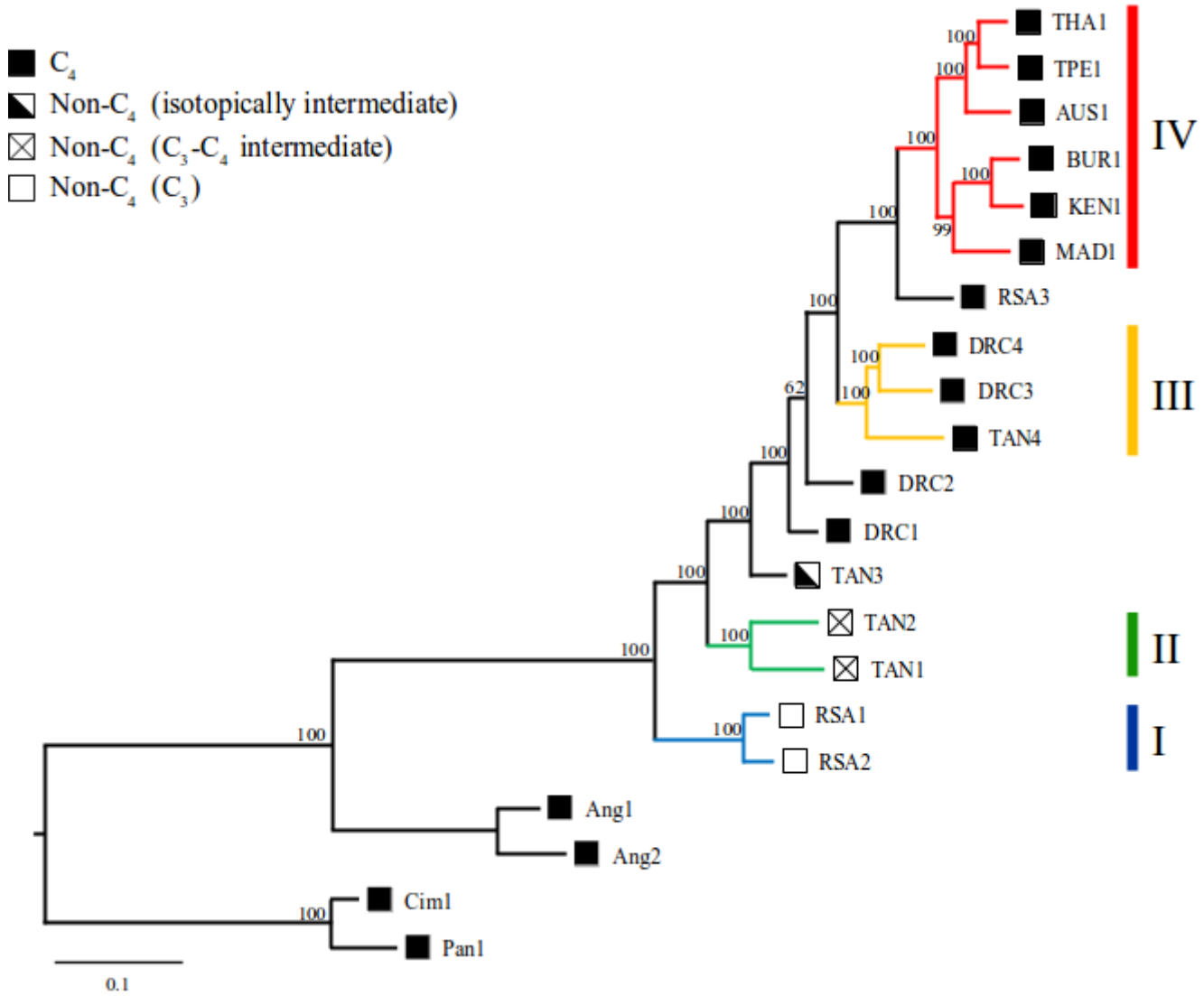
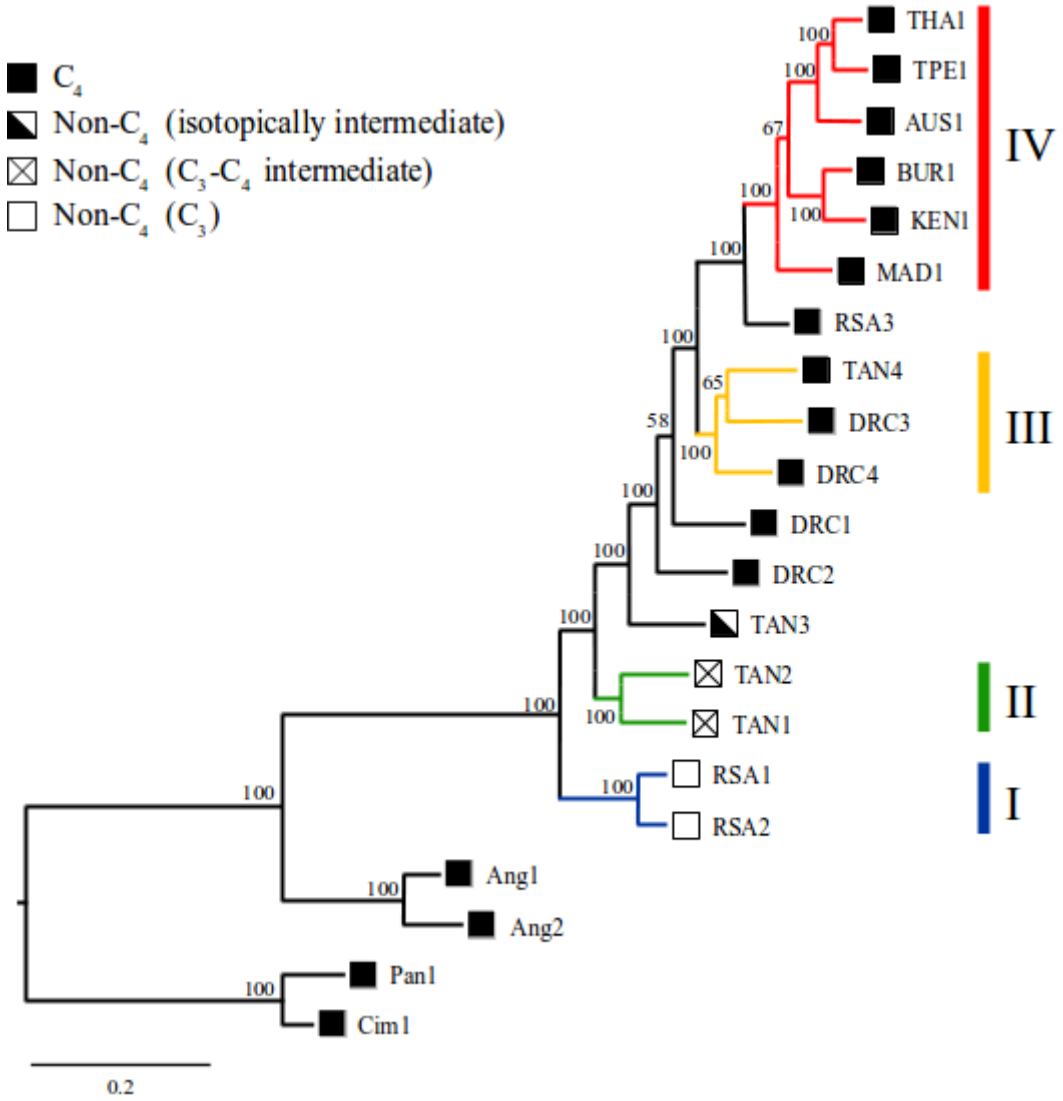

C4
Non-C4 (isotopically intermediate)
Non-C4 (C3-C4 intermediate)
Non-C4 (C3)
THA1
TPE1
AUS1
BUR1
KEN1
MAD1
RSA3
DRC4
DRC3
TAN4
DRC2
DRC1
TAN3
TAN2
TAN1
RSA1
RSA2
Ang1
Ang2
Cim1
Pan1
IV
III
II
I
0.1

Fig. 4.S4. Phylogenetic relationships based on a subset of the whole genome sequencing of *Alloteropsis* accessions.


C4
Non-C4 (isotopically intermediate)
Non-C4 (C3-C4 intermediate)
Non-C4 (C3)
THA1
TPE1
AUS1
BUR1
KEN1
MAD1
RSA3
TAN4
DRC3
DRC4
DRC1
DRC2
TAN3
TAN2
TAN1
RSA1
RSA2
Ang1
Ang2
Pan1
Cim1
IV
III
II
I
0.2

183

Fig. 4.S5. Assignment of *Alloteropsis semialata* individuals to genetic clusters based on a subset of the aligned reads from the whole genome sequencing.

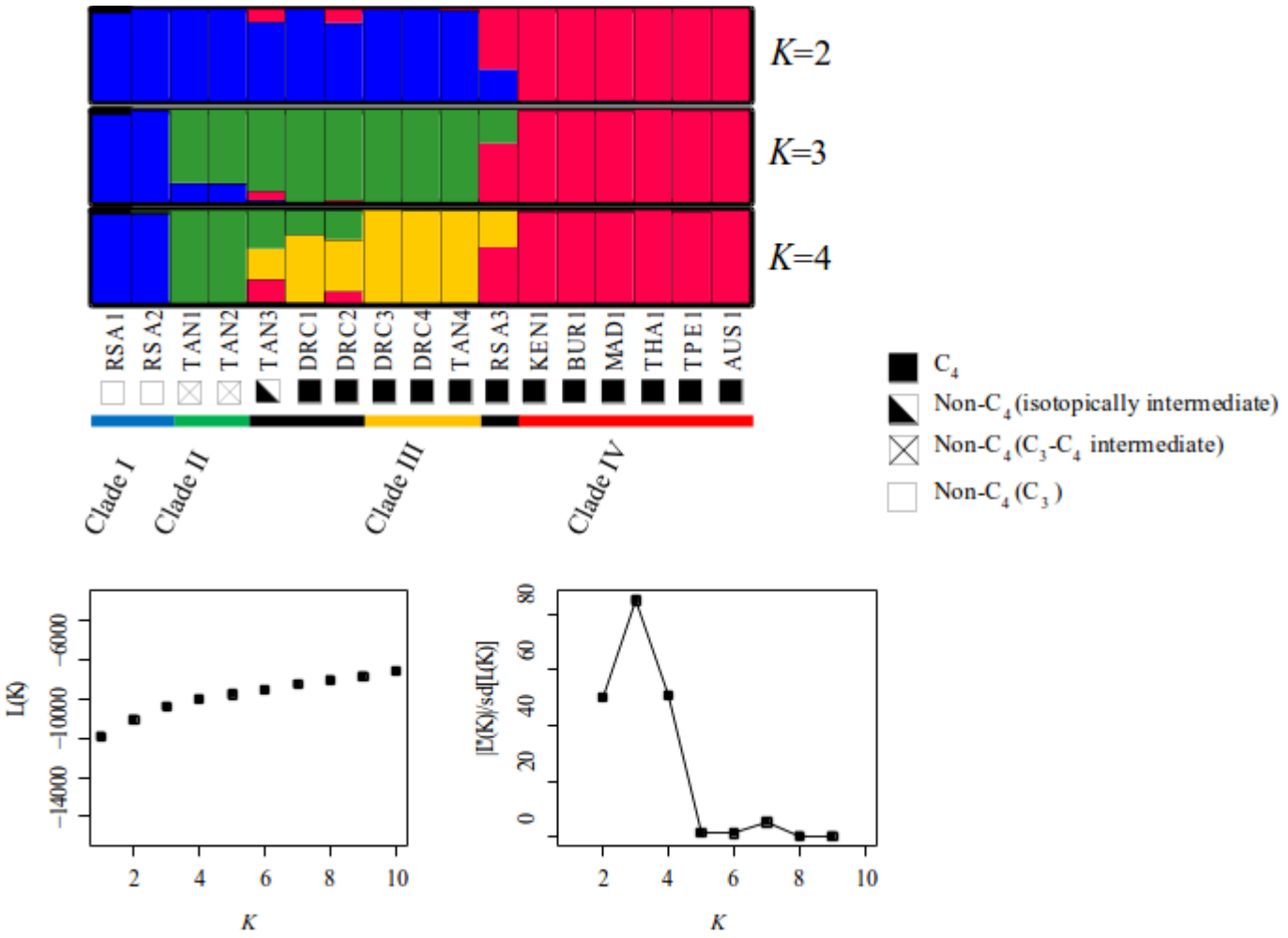

K=2
K=3
K=4
RSA1
RSA2
TAN1
TAN2
TAN3
DRC1
DRC2
DRC3
DRC4
TAN4
RSA3
KEN1
BUR1
MAD1
THA1
TPE1
AUS1
Clade I
Clade II
Clade III
Clade IV
C4
Non-C4 (isotopically intermediate)
Non-C4 (C3-C4 intermediate)
Non-C4 (C3)
L(K)
−14000
−10000
−6000
2
4
6
8
10
K
|L'(K)|/sd[L(K)]
0
20
40
60
80
2
4
6
8
10
K

Fig. 4.S6. Percentage of heterozygous sites for each accession.

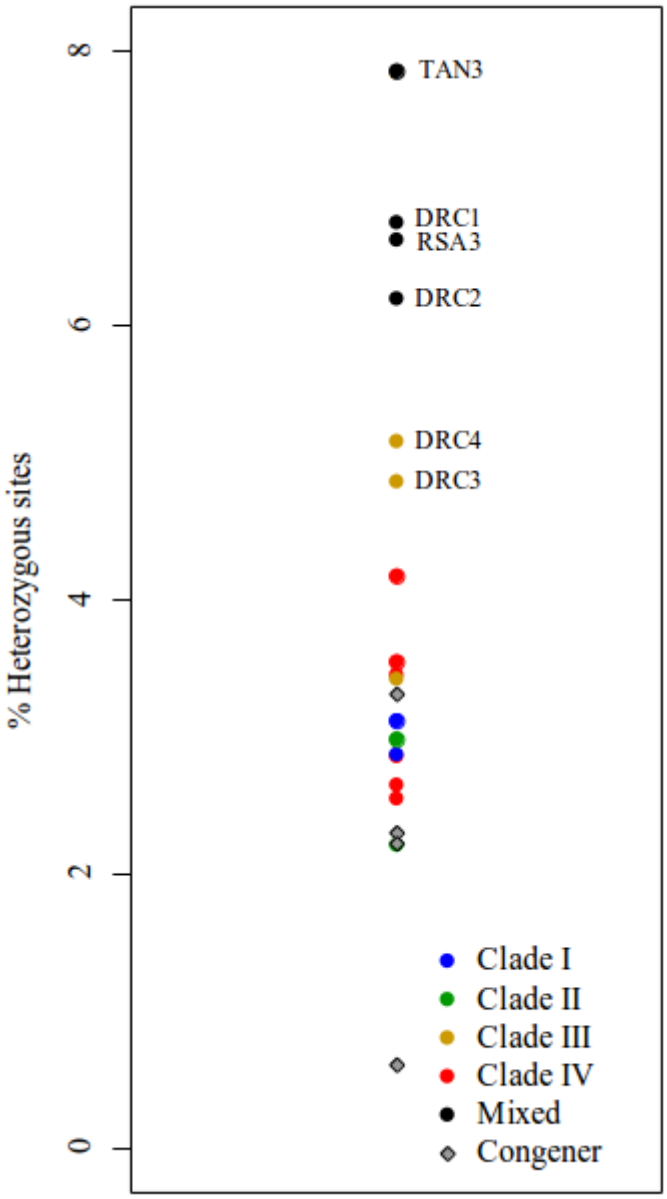

% Heterozygous sites
0
2
4
6
8
TAN3
DRC1
RSA3
DRC2
DRC4
DRC3
Clade I
Clade II
Clade III
Clade IV
Mixed
Congener

Fig. 4.S7. Results of PCR amplification of *ppc-IP3* and *pck-IP1* in *Alloteropsis*.

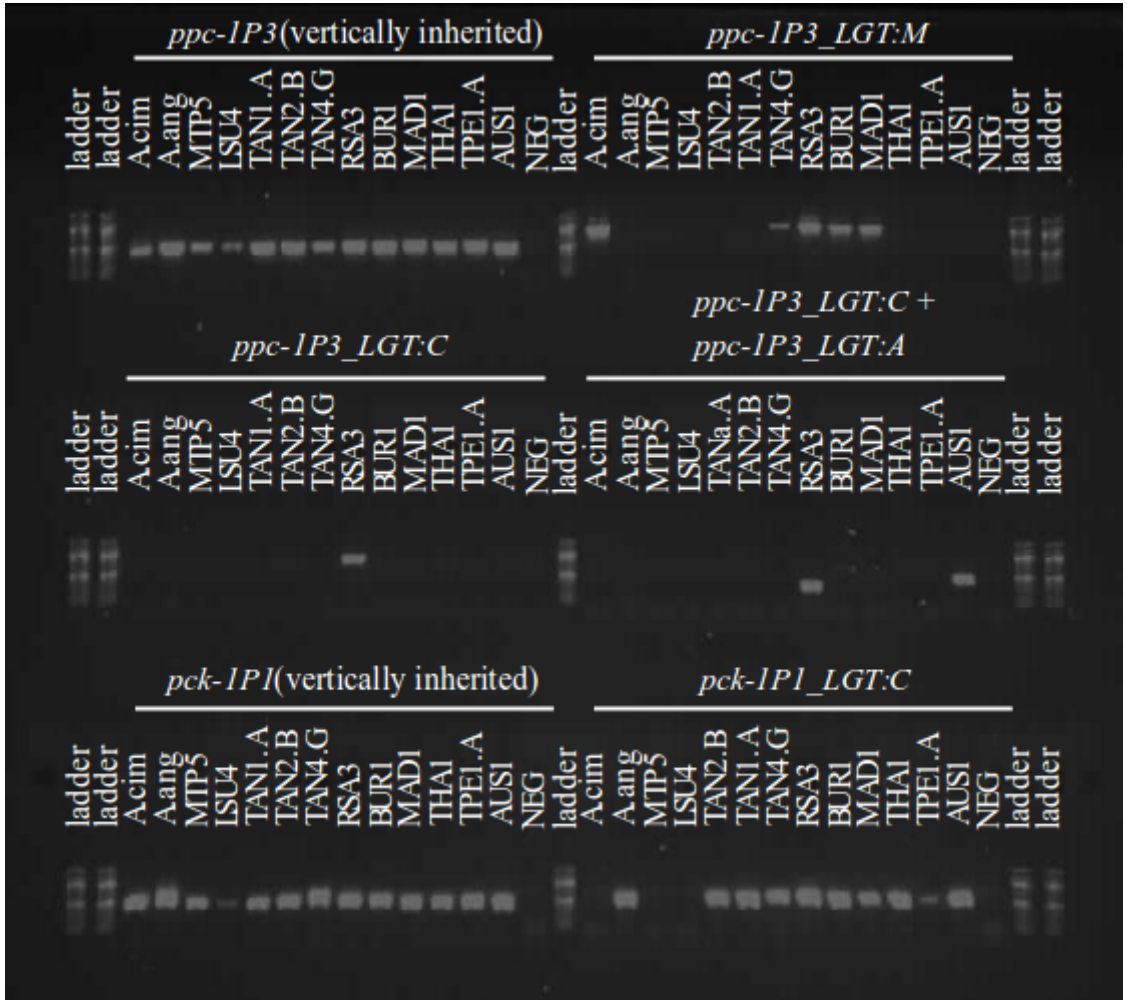

ppc-1P3 (vertically inherited)
ppc-1P3_LGT:M
ppc-1P3_LGT:C
ppc-1P3_LGT:C +
ppc-1P3_LGT:A
pck-1P1 (vertically inherited)
pck-1P1_LGT:C
ladder
ladder
Acim
A.ang
MTP5
LSU4
TAN1..A
TAN2.B
TAN4.G
RSA3
BUR1
MAD1
THA1
TPE1..A
AUS1
NEG
ladder
ladder

Fig. 4.S8. Phylogeny of *pck-1P1* in *Alloteropsis*.

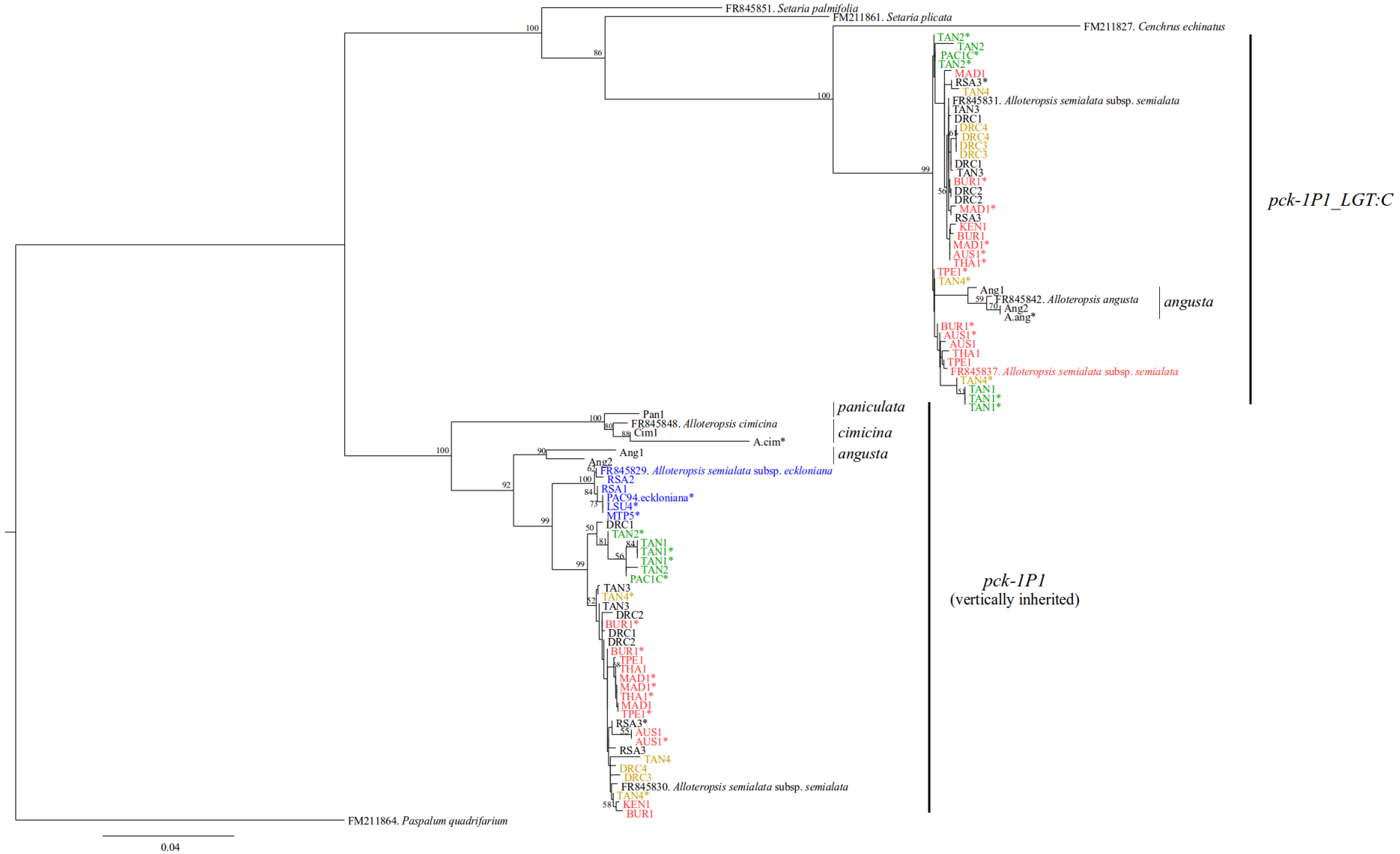

FR845851. Setaria palmifolia
FM211861. Setaria plicata
FM211827. Cenchrus echinatus
TAN2*
TAN2
PAC1C*
TAN2*
MAD1
RSA3*
TAN4
FR845831. Alloteropsis semialata subsp. semialata
TAN3
DRC1
DRC4
DRC4
DRC3
DRC3
DRC1
TAN3
BUR1*
DRC2
DRC2
MAD1*
RSA3
KEN1
BUR1
MAD1*
AUS1*
THA1*
TPE1*
TAN4*
Ang1
FR845842. Alloteropsis angusta
Ang2
A.ang*
BUR1*
AUS1*
AUS1
THA1
TPE1
FR845837. Alloteropsis semialata subsp. semialata
TAN4*
TAN1
TAN1*
TAN1*
pck-1P1_LGT:C
angusta
Pan1
FR845848. Alloteropsis cimicina
Cim1
A.cim*
Ang1
Ang2
FR845829. Alloteropsis semialata subsp. eckloniana
RSA2
RSA1
PAC94.eckloniana*
LSU4*
MTP5*
DRC1
TAN2*
TAN1
TAN1*
TAN1*
TAN2
PAC1C*
TAN3
TAN4*
TAN3
DRC2
BUR1*
DRC1
DRC2
BUR1*
TPE1
THA1
MAD1*
MAD1*
THA1*
MAD1
TPE1*
RSA3*
AUS1
AUS1*
RSA3
TAN4
DRC4
DRC3
FR845830. Alloteropsis semialata subsp. semialata
TAN4*
KEN1
BUR1
paniculata
cimicina
angusta
pck-1P1
(vertically inherited)
FM211864. Paspalum quadrifarium
0.04

Fig. 4.S9. Divergence times for different nodes estimated from vertically-inherited and laterally acquired genes.

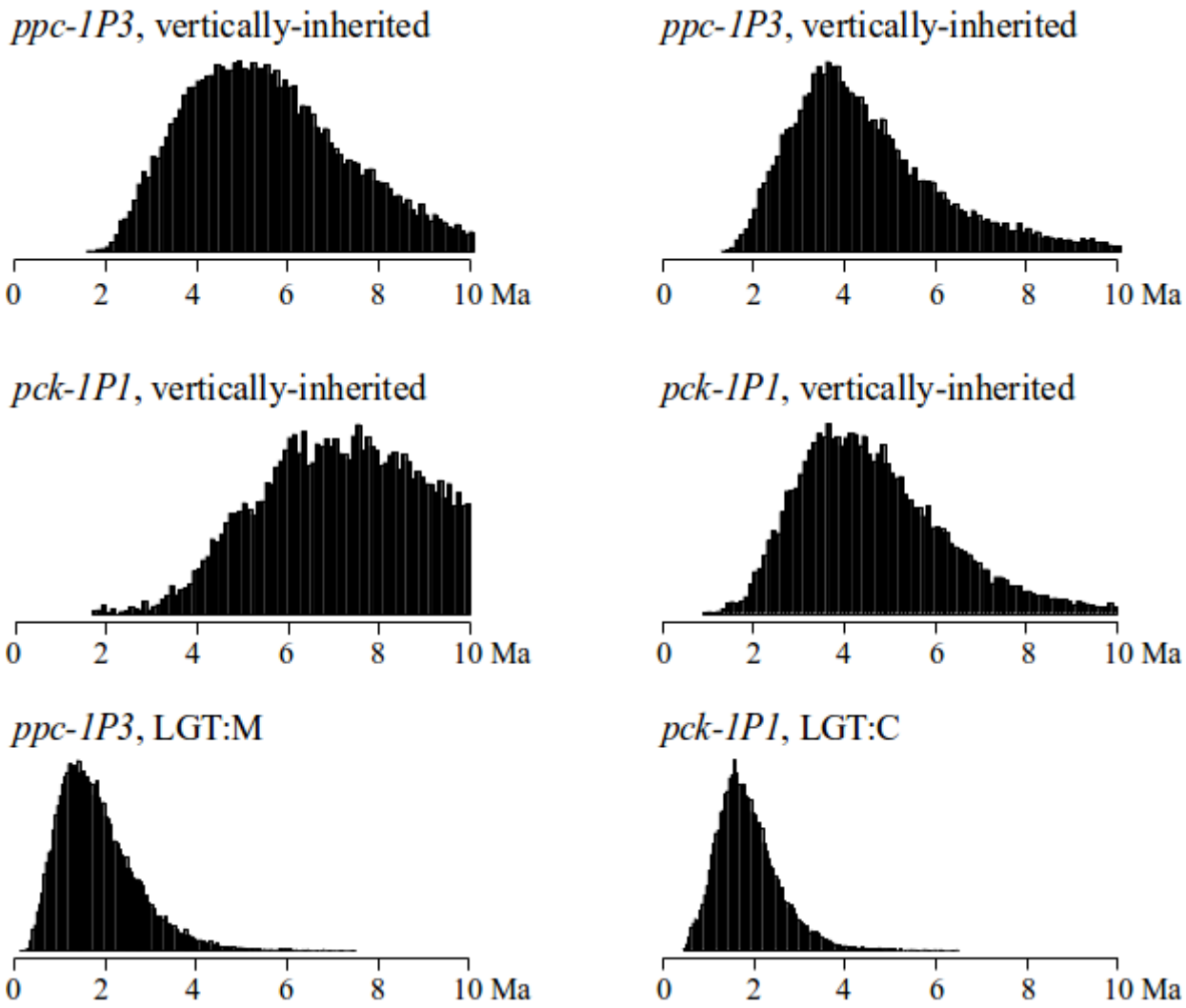

ppc-1P3, vertically-inherited
0 2 4 6 8 10 Ma
ppc-1P3, vertically-inherited
0 2 4 6 8 10 Ma
pck-1P1, vertically-inherited
0 2 4 6 8 10 Ma
pck-1P1, vertically-inherited
0 2 4 6 8 10 Ma
ppc-1P3, LGT:M
0 2 4 6 8 10 Ma
pck-1P1, LGT:C
0 2 4 6 8 10 Ma

Fig. 4.S10. Phylogeny of *ppc-1P3* in *Alloteropsis*.

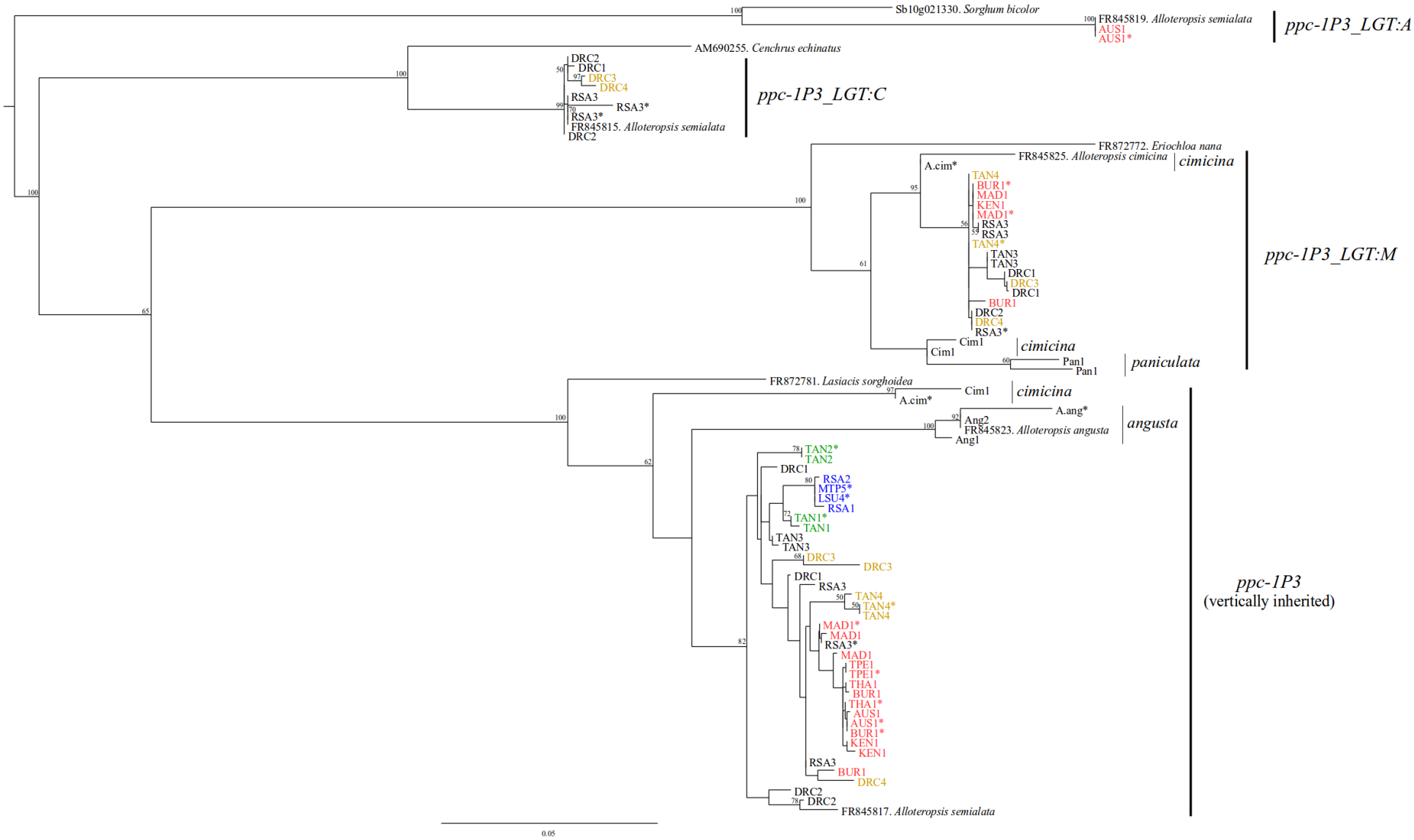

Sb10g021330. Sorghum bicolor
FR845819. Alloteropsis semialata
AUS1
AUS1*
ppc-1P3_LGT:A
AM690255. Cenchrus echinatus
DRC2
DRC1
DRC3
DRC4
RSA3
RSA3*
RSA3*
FR845815. Alloteropsis semialata
DRC2
ppc-1P3_LGT:C
FR872772. Eriochloa nana
FR845825. Alloteropsis cimicina
cimicina
A.cim*
TAN4
BUR1*
MAD1
KEN1
MAD1*
RSA3
RSA3
TAN4*
TAN3
TAN3
DRC1
DRC3
DRC1
BUR1
DRC2
DRC4
RSA3*
Cim1
Cim1
cimicina
Pan1
Pan1
paniculata
ppc-1P3_LGT:M
FR872781. Lasiacis sorghoidea
Cim1
cimicina
A.cim*
A.ang*
Ang2
FR845823. Alloteropsis angusta
angusta
Ang1
TAN2*
TAN2
DRC1
RSA2
MTP5*
LSU4*
RSA1
TAN1*
TAN1
TAN3
TAN3
DRC3
DRC3
DRC1
RSA3
TAN4
TAN4*
TAN4
MAD1*
MAD1
RSA3*
MAD1
TPE1
TPE1*
THA1
BUR1
THA1*
AUS1
AUS1*
BUR1*
KEN1
KEN1
RSA3
BUR1
DRC4
DRC2
DRC2
FR845817. Alloteropsis semialata
ppc-1P3
(vertically inherited)
0.05

Table 4.S1. Sample and sequencing information.

ID	Sample	Species	Genome size (Gb/1Cx²)/ploidy	PT[3] [$\delta^{13}C$]	Country	Voucher	Year	Latitude	Longitude	Sequencer	Read Length	Batch
Cim1	RCH20	*A. cimicina*	-	C_4 [-]	Madagascar	Hall 20 (K)	2011	-18.77	46.87	HiSeq 2500	100	2
Pan1[1]	MSV627	*A. paniculata*	-	C_4 [-]	Madagascar	Vorontsova 627 (K)	2011	-18.77	46.87	HiSeq 2500	101	6
Ang1[1]	Ang1	*A. angusta*	-	C_4 [-10.20]	DRC	Pauwels 1182 (BR)	1959	-4.04	21.76	HiSeq 3000	150	5
Ang2	3C	*A. angusta*	0.97/2n	C_4 [-]	Uganda	Namaganda & Wanyana 3C (MHU)	2009	-0.36	31.87	HiSeq 2500	100	2
RSA1	BL	*A. semialata*	-	Non-C_4 [-28.0]	South Africa	Lundgren & Ripley 11 (SHD)	2012	-29.71	29.96	HiSeq 2500	100	1
RSA2	JM	*A. semialata*	0.90/2n	Non-C_4 [-26.8]	South Africa	Ripley 1 (SHD)	2012	-33.32	26.44	HiSeq 2500	100	1
TAN1	L04C	*A. semialata*	1.10/2n	Non-C_4 [-23.1]	Tanzania	Lundgren & Christin 4 (SHD)	2014	-8.51	35.17	HiSeq 2500	100	2
TAN2	L01A	*A. semialata*	0.94/2n	Non-C_4 [-26.3]	Tanzania	Lundgren & Christin 1 (SHD)	2014	-5.63	32.69	HiSeq 2500	100	2
TAN3[1]	39688	*A. semialata*	-	Non-C_4 [-18.6]	Tanzania	Ruffo & Kisena 2806 (K)	1987	-7.87	31.67	HiSeq 2500	125	3
DRC1[1]	Asem4	*A. semialata*	-	C_4 [-10.7]	DRC	Kisimba & Malaisse 438 (BR)	2006	-10.36	26.08	HiSeq 3000	150	5
DRC2[1]	Asem2	*A. semialata*	-	C_4 [-11.2]	DRC	Lefebwe et al. 84 (BR)	1973	-10.42	26.18	HiSeq 3000	150	4
DRC3[1]	31768	*A. semialata*	-	C_4 [-10.6]	DRC	Poelman 92 (K)	1961	-11.64	27.48	HiSeq 2500	125	3
DRC4[1]	Asem3	*A. semialata*	-	C_4 [-10.7]	DRC	Bulaimu 743 (BR)	1973	-11.64	27.48	HiSeq 3000	150	5
TAN4	L02O	*A. semialata*	1.01/2n	C_4 [-11.4]	Tanzania	Lundgren & Christin 2 (SHD)	2014	-9.04	32.48	HiSeq 2500	100	2
RSA3	MD	*A. semialata*	0.87/6n	C_4 [-12.7]	South Africa	Ibrahim 20 (SHD)	2004	-25.76	29.47	HiSeq 2500	100	1
KEN1[1]	AB3722	*A. semialata*	-	C_4 [-11.7]	Kenya	Bogdan 3722 (EA)	1953	-0.02	37.91	HiSeq 2500	125	3
BUR1	Bur	*A. semialata*	0.98/2n	C_4 [-11.3]	Burkina Faso	Sanou BUR-734	2009	10.85	-4.82	HiSeq 2500	100	1
MAD1	Ma	*A. semialata*	1.03/2n	C_4 [-11.8]	Madagascar	Vorontsova 919 (K)	2013	-15.67	46.37	HiSeq 2500	100	1
THA1	ATSS837	*A. semialata*	-	C_4 [-12.2]	Thailand	AT & SS 837 (TCD)	2007	18.41	100.33	HiSeq 2500	100	2
TPE1	TW3	*A. semialata*	0.94/2n	C_4 [-14.6]	Taiwan	-	2014	24.47	120.72	HiSeq 2500	100	2
AUS1	Aus	*A. semialata*	1.1/2n	C_4 [-12.1]	Australia	AusTRCF 322458 0167	2005	-19.62	146.96	HiSeq 2500	100	1

[1] Newly sequenced sample; [2] 1Cx: monoploid genome size (DNA content per basic chromosome set; Greilhuber et al. 2005); [3]PT – photosynthetic type; [4] based on leaf anatomy (Christin *et al*. 2013b); [5]from Lundgren *et al.* 2015; [6]measured on a different accession from the same population; [7] Samples with the same number were sequenced together.

Table 4.S2. Alignment statistics of the *Alloteropsis* genome-skimming data to the *Setaria* reference genome.

Sample	Filtered pair-end reads	Theoretical coverage[1]	Total number of reads aligned	Pair-end reads concordantly aligned exactly one time	Total number of reads aligned to CDS	Individual coverage cut-off	Positions genotyped (% missing data)[2]	Positions genotyped in sub-set (% missing data)[3]
Cim1	20,415,006	1.86	1,176,915 (5.76%)	425,408 (2.08%)	438,307 (2.15%)	5	144,430 (15.4)	16,758 (26.6)
Pan1	7,740,126	0.70	792,297 (10.23%)	197,804 (2.56%)	227,351 (2.94%)	2	49,255 (71.1)	8,725 (61.7)
Ang1	14,343,018	1.96	579,842 (4.04%)	246,072 (1.72%)	175,529 (1.22%)	5	103,946 (39,1)	13,132 (42.4)
Ang2	18,067,598	1.86	831,340 (4.60%)	266,482 (1.47%)	295,435 (1.64%)	5	98,565 (42.2)	13,010 (43.0)
RSA1	14,326,452	1.30	751,861 (5.25%)	182,288 (1.27%)	247,120 (1.72%)	4	92,017 (46.1)	13,354 (41.4)
RSA2	12,069,794	1.34	523,241 (4.34%)	125,338 (1.04%)	169,927 (1.41%)	3	78,975 (53.7)	13,558 (40.6)
TAN1	18,821,504	1.71	1,404,829 (7.46%)	476,264 (2.53%)	397,122 (2.11%)	5	120,329 (29.5)	12,086 (47.0)
TAN2	20,041,562	2.13	1,532,343 (7.65%)	510,090 (2.55%)	412,690 (2.06%)	5	120,289 (29.5)	11,208 (50.9)
TAN3	33,537,244	3.81	1,856,861 (5.54%)	798,550 (2.38%)	570,630 (1.70%)	10	149,328 (12.5)	14,592 (36.0)
DRC1	33,118,910	4.52	1,997,844 (6.03%)	754,382 (2.28%)	590,736 (1.78%)	12	155,930 (8.6)	14,210 (37.8)
DRC2	23,353,806	3.18	1,291,044 (5.96%)	536,120 (2.30%)	404,876 (1.73%)	9	145,775 (14.6)	14,193 (37.8)
DRC3	28,530,034	3.24	2,595,576 (9.10%)	1,078,654 (3.73%)	763,728 (2.68%)	9	139,147 (18.5)	9,446 (58.6)
DRC4	14,480,696	1.97	908,170 (6.27%)	364,370 (2.52%)	260,236 (1.80%)	5	124,770 (26.9)	12,945 (43.3)
TAN4	18,395,178	1.82	1,233,515 (6.71%)	410,334 (2.23%)	349,951 (1.90%)	5	120,231 (29.5)	12,955 (43.2)
RSA3	13,396,464	1.54	611,585 (4.57%)	167,612 (1.25%)	217,359 (1.62%)	3	100,694 (41.0)	16,082 (29.5)
KEN1	24,717,950	2.80	1,703,210 (6.89%)	663,054 (2.68%)	418,286 (1.69%)	8	104,045 (39.0)	8,683 (62.0)
BUR1	13,103,476	1.33	578,025 (4.41%)	148,298 (1.13%)	173,457 (1.36%)	3	93,650 (45.1)	16,114 (29.4)
MAD1	16,120,906	1.57	836,086 (5.19%)	229,338 (1.42%)	245,711 (1.52%)	4	102,932 (39.7)	14,360 (37.1)
THA1	16,557,636	1.50	1,010,043 (6.10%)	313,698 (1.89%)	279,155 (1.69%)	4	118,067 (30.8)	14,089 (38.3)
TPE1	15,505,844	1.65	1,228,659 (7.92%)	423,360 (2.73%)	345,175 (2.23%)	4	109,410 (35.9)	11,449 (49.9)
AUS1	11,246,150	1.02	473,380 (4.21%)	118,012 (1.05%)	149,553 (1.33%)	3	79,564 (53.4)	14,446 (36.7)

[1] Estimated based on a genome sizes given in Table 4.S1; for accessions with unknown genomes size the largest value (1.1 Gb) was used; [2] Total of 170,629 across all accessions; [3] Total of 22,821 across all accessions

Table 4.S3. Primer pairs for amplification of genes copies of phosphoenolpyruvate carboxylase (*ppc*) and phosphoenolpyruvate carboxykinase (*pck*).

Gene copy	Forward primer	Reverse primer	Annealing temperature (ºC)	Extension time (s)
ppc-1P3_native	5'-GCTTCCGCACGCTGCAGCGG-3'	5'-CTCTGAGCACCTGGATGTTCC-3'	57	60
ppc-1P3_LGT:M	5'-AGCGTGAGTGCAAAGTGGCAG-3'	5'-GTGACCCTGAARAAKGGCCAC-3'	57	60
ppc-1P3_LGT:C	5'-GCGAGTGCCACATAAAGGAG-3'	5'-GTGACCCTGAARAAKGGCCAC-3'	57	60
ppc-1P3_LGT:A +ppc-1P3_LGT:C	5'-CGCTCCGTGGTCGTSAAGG-3'	5'-CAGGGTGACCCTGAAGAATG-3'	54	30
pck-1P1_native	5'-TGTCGACGGATCACAATAGGC-3'	5'-TACTCGATCGGGTACGCAGCC-3'	57	60
pck-1P1_LGT:C	5'-GACGACGCTGTCGACGGATCC-3'	5'-ACGGGTGTTCTCTGCATGCAG-3'	57	60

Chapter 5.

Tracking the origin and intraspecific spread of a laterally acquired gene involved in C_4 photosynthesis

Chapter 5. Tracking the origin and intraspecific spread of a laterally acquired gene involved in C_4 photosynthesis

Matheus E. Bianconi and Pascal-Antoine Christin

Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, UK

Personal contribution: I generated the genetic data, performed the genomic analyses and wrote the manuscript.

5.1. Abstract

A major question in evolutionary genetics is whether novel traits evolve from standing genetic variation or novel mutations. Addressing this problem in natural populations requires establishing when variants first appeared, which is complicated for point mutations. Recent lateral gene transfers (LGTs) provide tractable systems because their acquisition corresponds to a major novel mutation for the recipient species. Here we capitalize on a recent LGT of a gene for a key component of C_4 photosynthesis in the grass *Alloteropsis* to test the hypothesis that novel adaptive mutations can rapidly spread across established populations. Through comparisons of gene sequences of geographically spread individuals of the donor lineage, we identify the members of *Setaria palmifolia* complex in the Zambezian region of Africa as the likely origin of the LGT. We then screen whole genomes of multiple *A. semialata* accessions, and confirm that the LGT is restricted to some African individuals of this species with distinct genomic backgrounds, suggesting a rapid spread of the gene after the transfer. Using a new draft genome for one individual possessing the LGT, we show that the same scaffold includes another laterally acquired gene, but the two are separated by a vertically-inherited region. While the two LGTs are highly conserved among *A. semialata* accessions, the high levels of variation that exist in the flanking regions of vertically-inherited DNA do not support a rapid sweep via introgression. Instead, we propose that the genes spread via repeated integration in different parts of the genome, through mechanisms that remain to be elucidated.

Keywords: introgression, selective sweep, lateral gene transfer, C_4 photosynthesis, grasses.

5.2. Introduction

Determining the origin of an adaptive mutation and how it spreads through natural populations often requires large sample sizes and extensive genomic resources. Instances of recent lateral gene transfers (LGTs) can provide tractable systems to follow the fate of novel mutations, as these loci are easily identified, and their recombination with other, vertically-inherited variants is unlikely. Recent studies have reported several cases of LGT between multicellular eukaryotes (Li et al. 2014; Mahelka et al. 2017; Metzger et al. 2018). In plants, one of the most striking examples is the transfer to the grass *Alloteropsis* of multiple genes with a role in the complex C_4 photosynthetic pathway (Christin et al. 2012a; Chapter 4). The LGT instance in *Alloteropsis* provides an opportunity to investigate the fate of novel adaptive mutations associated with the evolution of complex traits.

The grass genus *Alloteropsis* comprises five species, four of which are C_4 (*A. angusta*, *A. cimicina*, *A. paniculata* and *A. papillosa*), and one, *A. semialata*, encompasses C_4 and non-C_4 populations (Ellis 1971; Lundgren et al. 2015). Multiple genes with a key role in the C_4 biochemical cycle were laterally acquired by *Alloteropsis* from other grass species (Christin et al. 2012a; Dunning et al. in prep; Chapter 4). Such genes encode enzymes that were optimized for the C_4 function in the donor species (Christin et al. 2007, 2012a), and they replaced the vertically inherited orthologs, which lack C_4-adaptive mutations (Christin et al. 2012a; Dunning et al. 2017; Chapter 4).

Out of the four laterally acquired genes recruited for C_4 photosynthesis in *Alloteropsis*, three encode the enzyme phosphoenolpyruvate carboxylase (PEPC), and were transferred from members of the grass lineages *Themeda*, *Setaria palmifolia* species complex, and Melinidinae in various regions of the world (Christin et al. 2012a; Chapter 4). The *Themeda* LGT is restricted to Australia and parts of Asia (Dunning et al. in prep), while the *Setaria* and Melinidinae genes occur in Africa (Chapter 4). The fourth gene encodes the enzyme phosphoenolpyruvate carboxykinase (PCK), and was acquired from a member of *Cenchrus* (Christin et al. 2012a). Although previous studies have established the identity of the donor species in terms of taxonomic groups, in some cases at the level of the genus or group of species, a comprehensive investigation of populations of putative donors growing in proximity to *Alloteropsis* was not conducted,

so that their identity cannot be precisely pinpointed to species or lineages within each species. Dedicated screens of both recipient and donor lineages are therefore required.

The complex history of spread of LGTs among *Alloteropsis* populations has shed light on the diversification of the group and on the evolution of C_4 photosynthesis. Out of the four LGTs, two were secondarily spread among *A. semialata* and *A. angusta* (PCK), and *A. semialata* and *A. cimicina* (PEPC Melinidinae; Dunning et al. 2017). Within *A. semialata*, laterally acquired genes were spread within or among distinct C_4 lineages, and also introgressed into non-C_4 individuals through admixture events between C_4 and non-C_4 lineages (Chapter 4). However, the geographical range of the occurrence of LGTs is not clear, particularly within Africa, where two distinct laterally acquired genes for PEPC are present, and sometimes coexist within the same individual (Christin et al. 2012a; Chapter 4). Evidence of pseudogeneization of one of the genes in some individuals suggests that the two LGTs can replace each other over time (Chapter 4). The dynamics controlling the spread of these LGTs however still remain poorly understood, mainly because previous studies included small sample sizes.

In this study, we conduct genomic analyses in a spatial context to elucidate the dynamics that dictated the fate of the PEPC gene after its lateral acquisition from a member of the *Setaria palmifolia* complex. We first compare the PEPC gene sequences of geographically distant individuals of the *Setaria palmifolia* complex to (i) identify the most likely donors, in terms of intraspecific lineages. A diversity of *A. semialata* individuals collected across Africa is then screened to (ii) determine the distribution of the LGT within the recipient species. We then generate and assemble a genome model for one individual possessing the LGT to (iii) establish the size and content of the laterally acquired fragment. The genomic region containing the LGT is then compared across 24 individuals of *A. semialata* from across the species range to (iv) investigate the hypothesis that selection drove the rapid sweep of the LGT throughout Africa via rapid introgression involving recombination with native chromosomes. The adopted biogeographic framework allows reconstructing the order of genomic exchanges leading to the fixation of a functional LGT within the recipient species.

5.3. Material and Methods

5.3.1. Analyses of the LGT donor lineage

A total of 23 accessions of the grass subtribe that encompasses the *Setaria palmifolia* complex (i.e. Cenchrinae) were analysed in this study (Table 5.S1). First, plastid and nuclear markers were isolated via PCR in order to establish the relationships among the accessions of the *S. palmifolia* complex, for which hybridization was suggested based on morphology (Clayton 1979). DNA was extracted from fresh or dry leaves using DNeasy Plant Mini kit (Qiagen). Grass-specific primers for the chloroplast marker *trnK-matK* were retrieved from Hilu et al. (1999). PCR reactions were prepared using 10-40 ng of gDNA template, 5 µl of 5X GoTaq Flexi reaction buffer (Promega, USA), 2 mM of $MgCl_2$, 0.08 mM of dNTPs, 0.2 µM of each primer and 0.5 unit of GoTaq polymerase (Promega) in a total volume of 25 µl. The PCR mixtures were initially incubated in a thermocycler for 2 min at 94°C followed by 35 cycles consisting of 30 s at 94°C for denaturation, 1 min at 54°C for primer annealing, and 1 min at 72°C for elongation, with an additional 10 min at 72°C for final elongation. Then, to identify the most likely donor species among the 23 accessions, the sequences of the phosphoenolpyruvate carboxylase gene laterally transferred to *Alloteropsis* (*ppc1P3*) were isolated via PCR. A fragment comprising exons 8 to 10 of *ppc1P3* was amplified using a pair of specific primers retrieved from Chapter 4. PCR reactions were set as described above. All successfully amplified *ppc1P3* and *trnK-matK* were then cleaned using Exo-SAP-IT (Affymetrix, Santa Clara, CA, USA), and Sanger-sequenced. Sequencing chromatograms were individually inspected and double peaks were called using IUPAC codes for ambiguous bases. A phylogenetic tree based on *trnk-matK* was inferred using MrBayes v.3.2.2 (Ronquist et al. 2012) with the GTR+G substitution model. Two analyses consisting of four chains each were performed in parallel for 10,000,000 generations. The burn-in period was set to 25%, and a consensus tree was inferred for each gene from all trees obtained post burn-in in the parallel analyses.

Genome-wide, low-coverage sequencing data of three individuals of the *S. palmifolia* complex were retrieved from another study (Park et al. in prep; Table 5.S1). These were selected because preliminary phylogenetic analyses suggested that they might be closely related to the donor lineage (see Results). Genomic reads for an additional Cenchrinae representative, *S. italica*, were retrieved from NCBI SRA database (refer to Chapter 2 for a list of pooled samples retrieved here).

5.3.2. Relationships among A. semialata

To track the spread of the genetic material that was laterally acquired by *A. semialata*, genomic data was obtained for accessions capturing the phylogenetic and geographic diversity of this group. First, low-coverage and resequencing genome-wide datasets of 49 individuals of *Alloteropsis*, two of sister lineages within the same panicoid grass tribe (*Panicum pygmaeum* and *Entolasia marginata*), and one of a different tribe (*Themeda triandra*) were retrieved from previous studies (Table 5.S1; Lundgren et al. 2015; Dunning et al. in prep; Chapter 3; Chapter 4). A second batch of whole genome datasets was generated for this study, with 30 accessions of *A. semialata*, *A. angusta* and *A. cimicina* sequenced at low coverage (Table 5.S1). For this, DNA was extracted either from fresh or dry leaves as described above. DNA libraries were prepared according to Lundgren et al. (2015), and samples were pooled before sequencing. Samples were sequenced at the Genopole platform of Toulouse. All genomic reads were filtered prior to analyses using NGSQC Toolkit v.2.3.3 (Patel and Jain 2012) to remove adapter contamination, and reads having less than 80% of the sites with Phred score > 20. Reads were then trimmed from the 3' end to remove low quality bases also using NGSQC Toolkit.

Phylogenetic analyses of plastomes and nuclear genomes were performed to infer the relationships among the accessions of *A. semialata*. Full plastomes were either retrieved from previous studies (Lundgren et al. 2015; Chapter 4) or assembled here using the genome datasets. For the latter, plastome sequences were retrieved from NCBI (accessions KT281146.1, NC_027952.1 and KT281153.1 of *A. angusta, A. cimicina and A. semialata*, respectively, and KJ001642.1 of *S. italica*) and used as references for mapping the reads of closely related individuals. Paired-end reads were mapped using Bowtie2 v.2.1.0 (Langmead and Salzberg 2012) using default parameters, and a majority consensus sequence was called for each individual using Geneious v.6.1.8 (Kearse et al. 2012). Sequences were then aligned using MAFFT v.7.13 (Katoh and Standley 2013), and the inverted repeat region was removed before the analysis to avoid using the same sequence twice in the alignment. A maximum likelihood (ML) tree was inferred using RAxML v.8.2.4 (Stamatakis 2014) with the GTR+CAT model of substitution, and node support was evaluated with 100 bootstrap pseudoreplicates.

A nuclear genome tree was inferred using gene sequences reconstructed using the genomic datasets. A similar approach to that used in Chapter 2 was adopted. Briefly,

a reference nuclear dataset was retrieved from Dunning et al. (2017), which consists of coding sequences (CDSs) of co-orthologous genes at the Panicoideae subfamily level. This dataset contains all genes derived from a single gene in the common ancestor of the group via speciation and gene duplication events that were present in the transcriptomes/genomes of *A. semialata*, *S. italica* and *Sorghum bicolor*. The final dataset used here consisted of CDSs of 11,313 genes, and only *A. semialata* sequences were used to avoid multiple references. Genomic data were mapped as single-end reads to this reference dataset using Bowtie2 with default parameters and the local alignment option. Consensus sequences were then called for each of the 11,313 genes using a pipeline modified from Chapter 4, and IUPAC codes for ambiguous bases were used for all polymorphic sites. Although considering all polymorphisms might incorporate sequencing errors in the final dataset, these would be randomly distributed and would therefore not systematically bias the phylogenetic analyses. All genes were then concatenated to generate a 6,138,916 bp supermatrix. This supermatrix was subsequently trimmed using trimAl v.1.4 (Capella-Gutiérrez et al. 2009) to remove sites with missing data in more than 10% of the accessions. A ML tree was then inferred on the 298,850 bp resulting supermatrix using RAxML as described above for the plastome dataset.

Genome sizes were either retrieved from previous studies (Lundgren et al. 2015; Chapter 4), or measured in this study using the method described in Chapter 4. Carbon isotope analyses were used to distinguish C_4 and non-C_4 individuals, and were retrieved from previous studies (Lundgren et al. 2015; Lundgren et al. in prep; Chapter 4).

5.3.3. Distribution of the laterally acquired PEPC gene within A. semialata

Another 81 accessions of *A. semialata* collected from several locations in Africa were screened for the presence of the PEPC gene laterally acquired from a member of the *S. palmifolia* complex (*ppc1P3_LGT:C*; Table 5.S1). To test for presence/absence of *ppc1P3_LGT:C*, DNA was extracted and PCR reactions were performed as described above for accessions of the *S. palmifolia* complex, a method used previously on a smaller sample size (Chapter 4). Positive and negative controls were included, which consisted of good quality DNA samples of *A. semialata* accessions known to carry and lack *ppc1P3_LGT:C*, respectively (Chapter 4). PCR products were then visualized using

gel electrophoresis, and amplified products were purified using Exo-SAP-IT and Sanger-sequenced to confirm the gene identity.

5.3.4. Genome sequencing and assembly

A draft genome of a C_4 individual of *A. semialata* from Zambia known to carry *ppc1P3_LGT:C* (based on the analyses described above) was assembled in this study, using a combination of short reads from Illumina paired-end sequencing and long reads from Pacbio sequencing. Genomic DNA was extracted from fresh leaves using the Dneasy Kit Maxi Kit (Qiagen) following the manufacturer instructions. Genomic libraries for short-read sequencing were prepared with an insert size of 550 bp and sequenced in a full lane of Ilumina HiSeq 3000. A total of 220,035,230 paired-end reads of 250 bp were generated (sequencing depth ~ 25 X). Both library preparation and sequencing were performed at the Edinburgh Genomics Centre. Long-read sequencing was performed on a SMRT PacBio platform at the Centre for Genomic Research at the University of Liverpool, and generated 6,169,470 filtered subreads with a mean length of 5,905 bp (sequencing depth ~ 16 X).

A draft genome was assembled using a hybrid strategy with Illumina and Pacbio data, following the approach used by Dunning et al. (in prep) to assemble the genome of an *A. semialata* individual from Australia. Illumina reads were initially filtered and trimmed using NGSQCToolkit as described above, and duplicate reads were removed using PrinSeq-lite v.0.20.3 (Schmeider and Edwards 2011). The remaining sequences were then error-corrected using SOAPec v.2.01 (Luo et al. 2012). The cleaned, error-corrected Illumina reads were then assembled into contigs using SOAPdenovo2 v.2.04 (Luo et al. 2012) with the parameters *k-mer* = 65, *KmerFreqCutoff* = 10, *mergeLevel* = 3 (max), and *arcWeight* = 5. Pacbio long reads were error-corrected using Proovread v.2.14.0 (Hackl et al. 2014) using the cleaned Illumina reads. The hybrid assembly was then performed using Dbg2olc v.11062016 (Ye et al. 2016) with the SOAPdenovo contigs and the error-corrected Pacbio reads. Different values of parameters *KmerCovTh* (2-3), *AdaptiveTh* (0.001 – 0.01), and *MinOverlap* (10-30) were used with a k-mer size of 17, generating 31 draft assemblies. The N50 ranged from 24.5 to 26.4 Kb (mean = 25.6 Kb), and the assembly length ranged from 818.9 Mb to 936.5 Mb (mean = 874.4 Mb), which corresponds to 75-86% of the genome as estimated by flow cytometry (Table 5.S2). The longest assembly was selected for analyses. The

completeness of the genome was assessed by checking for the presence of 956 plant benchmarking universal single-copy orthologs using BUSCO v.1.22 (Simão et al. 2015).

5.3.5. Analyses of the fragment containing the LGT

Genomic paired-end reads of the different accessions were individually mapped onto the draft genome of the Zambian *A. semialata* using Bowtie2 with default parameters. The scaffold containing *ppc1P3_LGT:C* was identified, and consensus sequences were called using the read alignments (see above) to reconstruct the whole fragment for each accession. Phylogenetic analyses were then performed on the whole genomic fragment, and on separate segments within the fragment. First, regions potentially acquired via LGT along with *ppc1P3_LGT:C* were identified after inspecting the BAM alignments of *S. palmifolia* complex representatives and *S. italica* using Geneious. Genomic segments covered by at least three reads of one of the *Setaria* accessions were considered putative laterally acquired fragments. However, only segments with high nucleotide identity (> 95%) between *Setaria* and the reference genome were used for subsequent phylogenetic analysis of LGT regions; low identity segments were considered ambiguous and not used for analysis (see Results). In addition to the genomic sequence of *ppc1P3_LGT:C*, these included (1) a ~ 160-bp segment 5' upstream to *ppc1P3_LGT:C*, (2) a ~ 19.3-Kb immediately 3' downstream to *ppc1P3_LGT:C,* (3) a ~ 17.5-Kb segment containing tandem duplicates of the gene Serine/Threonine PP1 (*S/T PP1*; see Results), also located 5’ upstream to *ppc1P3_LGT:C*. All other regions were similar to all *A. semialata* but not the *Setaria* samples, and were considered as non-laterally acquired. ML trees were inferred for laterally and non-laterally acquired segments using RAxML with a GTR+CAT model of substitution as described above. A first tree was inferred from the non-laterally acquired region separating the two LGTs to determine whether it was transmitted along with the LGTs across populations. Trees were subsequently inferred for the two LGTs and their flanking regions. Note that phylogenetic analyses of *ppc1P3_LGT:C* included all genomic datasets of *S. palmifolia* complex and *A. semialata* (see above).

5.4. Results

5.4.1. Putative LGT donor lineage

The relationships among the Cenchrinae accessions were investigated using a chloroplast (*trnK-matK*) marker, for which sequences for members of the group are publicly available. The *trnK-matK* tree shows that all samples sequenced here are nested in a highly supported group including members of *S. incrassata*, *S. verticillata*, *S. sphacelata* and *S. palmifolia* complex (e.g. *S. palmifolia* and *S. megaphylla*; Fig. 5.1). Among these, four accessions formed a highly supported group with the *S. palmifolia* complex members (SPC01, SPC03, SPC16 and SPC23), while another six accessions (SPC32, SPC33, SPC34, SPC36, SPC37 and SPC38) formed a group with *S. verticillata* and *S. sphacelata*. To identify the potential donor lineage, we inferred a phylogenetic tree using sequences of *ppc1P3* of accessions of Cenchrinae and sequences of the laterally acquired gene of *Alloteropsis* (*ppc1P3_LGT:C*). In this tree, a subset of 11 Cenchrinae accessions formed a highly supported group (bootstrap support, bs = 99) with *Cenchrus* and *Pennisetum* accessions (Fig. 5.2). All other Cenchrinae accessions were nested in a highly supported group (bs = 97), which consisted of sequences of members of the *S. palmifolia* complex and all *A. semialata* (Fig. 5.2). Within this second group, three accessions sequenced in this study (SPC01, SPC16 and SPC23) and one previously sequenced (AM690293; Christin et al. 2012) formed a monophyletic group with all *A. semialata* (bs = 53).

5.4.2.Phylogeographic distribution of the laterally acquired PEPC gene within A. semialata

Plastome and nuclear genome trees recovered all major lineages of *Alloteropsis* as reported in previous studies (Lundgren et al. 2015; Chapter 4), and confirmed the different evolutionary histories of plastomes and nuclear genomes within *A. semialata* (Fig. 5.3). Our screening of whole genome sequencing datasets confirmed that *ppc1P3_LGT:C* is restricted to African clades of *A. semialata* (Figs 5.3 and 5.4). This includes (1) most members of a subgroup of individuals from the Zambezian region (Zambia, DRC, Tanzania, Malawi and North of Mozambique) that form a well

supported clade in the nuclear tree (Clade III), (2) individuals including identified and putative polyploids from South Africa and South of Mozambique carrying the gene and placed between the major C_4 lineages of Africa and Southeastern Asia/Australia (Clades IVa and IVb), and (3) polyploid individuals from Cameroon, which are grouped together with known admixed individuals between the non-C_4 and the C_4 lineages from the Zambezian region that were previously reported to carry the gene (Figs 5.3 and 5.4; Table 5.S2; Chapter 4). Interestingly, we found two locations in Zambia (ZAM15-03 and ZAM15-07) and one in Tanzania (TAN16-03) where individuals with and without the gene coexist. The two populations from Zambia include non-C_4 and C_4 genes that belong to different clades, and the LGT is restricted to the C_4 individuals. By contrast, the Tanzanian population contains only C_4 individuals, and the individuals with and without the gene are very closely related on both nuclear and plastid phylogenies (Fig. 5.3). Visual inspection of read depth of resequencing datasets confirmed the intrapopulation variation in presence/absence of the LGT (Fig. 5.5). A larger screening conducted in the Tanzanian population showed that out of 17 individuals, four carry the gene (Table 5.S1).

5.4.3.Draft genome assembly

The draft genome of one individual of nuclear clade III from Zambia (ZAM15-05-10; Fig. 5.3a) was assembled in this study. The assembly generated here has a N50 of 24.8 Kb with 50,719 contigs (longest contig = 229,107 Kb), and is 936.5 Mb long, which is close to the expected genome size of this individual (1,081.5 Mb; Table 5.S2). The BUSCO analyses to assess the completeness of the assembly showed that out of the 956 single-copy genes analysed, 873 are complete single-copy, while 392 are duplicated, 32 are fragmented and 51 are missing. The percentage of genomic reads of other *A. semialata* accessions mapped to ZAM15-05-10 was 67.6% on average, ranging from 10.4% to 87.3%. For *A. angusta*, *A. cimicina* and Cenchrinae accessions the average percentages were 26.7%, 13.2% and 14.6%, respectively, which is expected since mapping among distant relatives works mainly for coding sequences (Chapter 4). All *ppc* paralogs identified by previous studies in *A. semialata* genomes were retrieved as complete sequences in this draft genome, including the laterally acquired copies from *Setaria* (*ppc1P3_LGT:C*) and Melinidinae (*ppc1P3_LGT:M*; Christin et al. 2012a). The scaffold containing *ppc1P3_LGT:C* is 106.3 kb long.

5.4.4. Analyses of the genomic fragment containing ppc1P3_LGT:C

Analyses of open reading frame (ORF) sequences in the genomic scaffold containing *ppc1P3_LGT:C* revealed another gene potentially acquired from Cenchrinae. BLAST analyses annotated this gene as a serine/threonin-protein phosphatase PP1 (*ST/PP1*). This gene is a tandem duplicate in the Zambian *A. semialata* genome and is situated 27.5 Kb 5' upstream to the transcription start site of *ppc1P3_LGT:C* (Fig. 5.6). In the complete genome of *S. italica*, *ST/PP1* is located 6.8 Kb apart from *ppc1P3*, on chromosome 4 (Fig. 5.6). The distribution of this gene in *A. semialata* individuals matches the distribution of *ppc1P3_LGT:C*, which suggests that the two were transferred across plants as part of the same LGT fragment. Phylogenetic analyses of the two copies show that these were duplicated after the divergence between *S. italica* and *S. palmifolia* species (Fig. 5.7). The coverage of *A. semialata* individuals carrying the LGT fragments is approximately constant in most parts of the whole scaffold, although there are regions with no coverage at all, suggesting the occurrence of multiple genomic rearrangements after the LGT fragment was acquired. Nonetheless, laterally transferred regions sum up to approximately 40% of the whole fragment. Phylogenetic trees using either the laterally acquired or the putatively vertically inherited segments (defined here as regions with no coverage in *Setaria* accessions) were then inferred to investigate the dynamics of the LGT fragment in *A. semialata*. The tree inferred from the vertically inherited region between the laterally acquired genes (*ppc1P3_LGT:C* and *ST/PP1*) recovers some groups of the nuclear genome and plastome trees of *A. semialata*, including the Asian/Australian and the South African clades (Fig. 5.8). Most accessions from Zambia and Tanzania form a weakly supported group, but the phylogenetic tree lacks overall resolution. Importantly, the accessions carrying the LGTs are not grouped together as would be expected if the segment had been rapidly spread along the LGTs. On the other hand, the tree inferred from the immediate 3' downstream segment of *ppc1P3_LGT:C* shows high sequence similarity among accessions carrying the LGT, with no groups clearly resolved (Fig. 5.9). The lack of variation among *A. semialata* accessions is compatible with a rapid spread of the segment, as seen on *ppc1P3_LGT:C* coding sequence (Fig. 5.2). Note that this laterally acquired segment is not present in the

accessions that do not have the LGT. Analysis of the 5' upstream region of *ppc1P3_LGT:C* shows a discontinuity in the read coverage at ~ 60 bp from the start codon in all accessions. Reads mapped to the first exon of *ppc1P3_LGT:C* do not have their mates in the 5' upstream region, which suggests that the discontinuity is longer than 550 bp (i.e. average insert size). However, the coverage is restored at ~ 130 bp 5' upstream and sequence analysis supports the homology between the reference genome sequence and those accessions carrying the laterally acquired fragments.

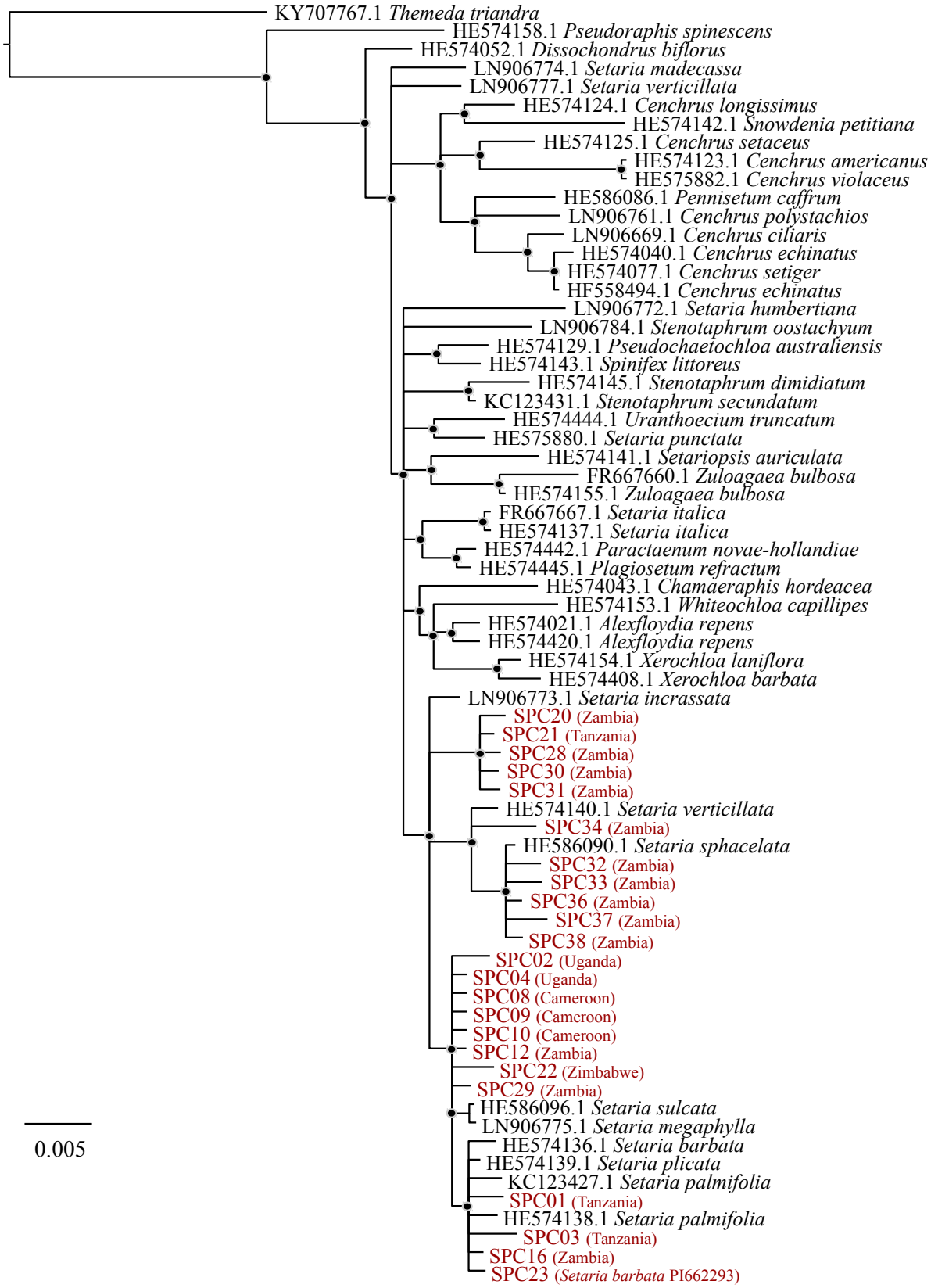


Fig. 5.1. Bayesian tree of Cenchrinae subtribe based on the chloroplast marker *trnK-matK*. Accessions analysed in this study are coloured in red. Black circles on nodes are posterior probabilities > 50% (values ≤ 50% were omitted).

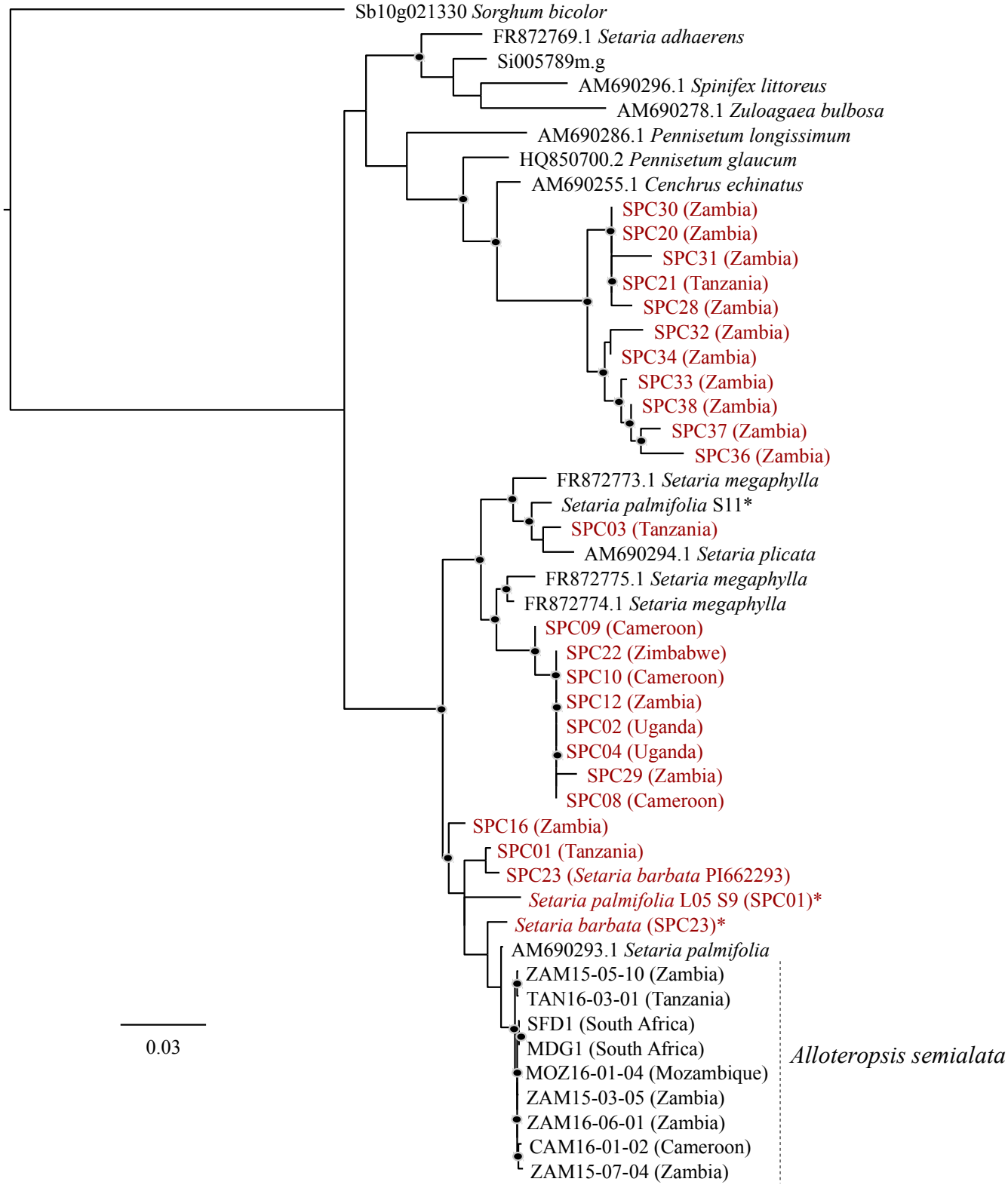


Fig. 5.2. Maximum likelihood tree of *ppc1P3_LGT:C*. Cenchrinae accessions analysed in this study are coloured in red. Black circles on nodes are bootstrap support values > 50 (values ≤ 50 were omitted).

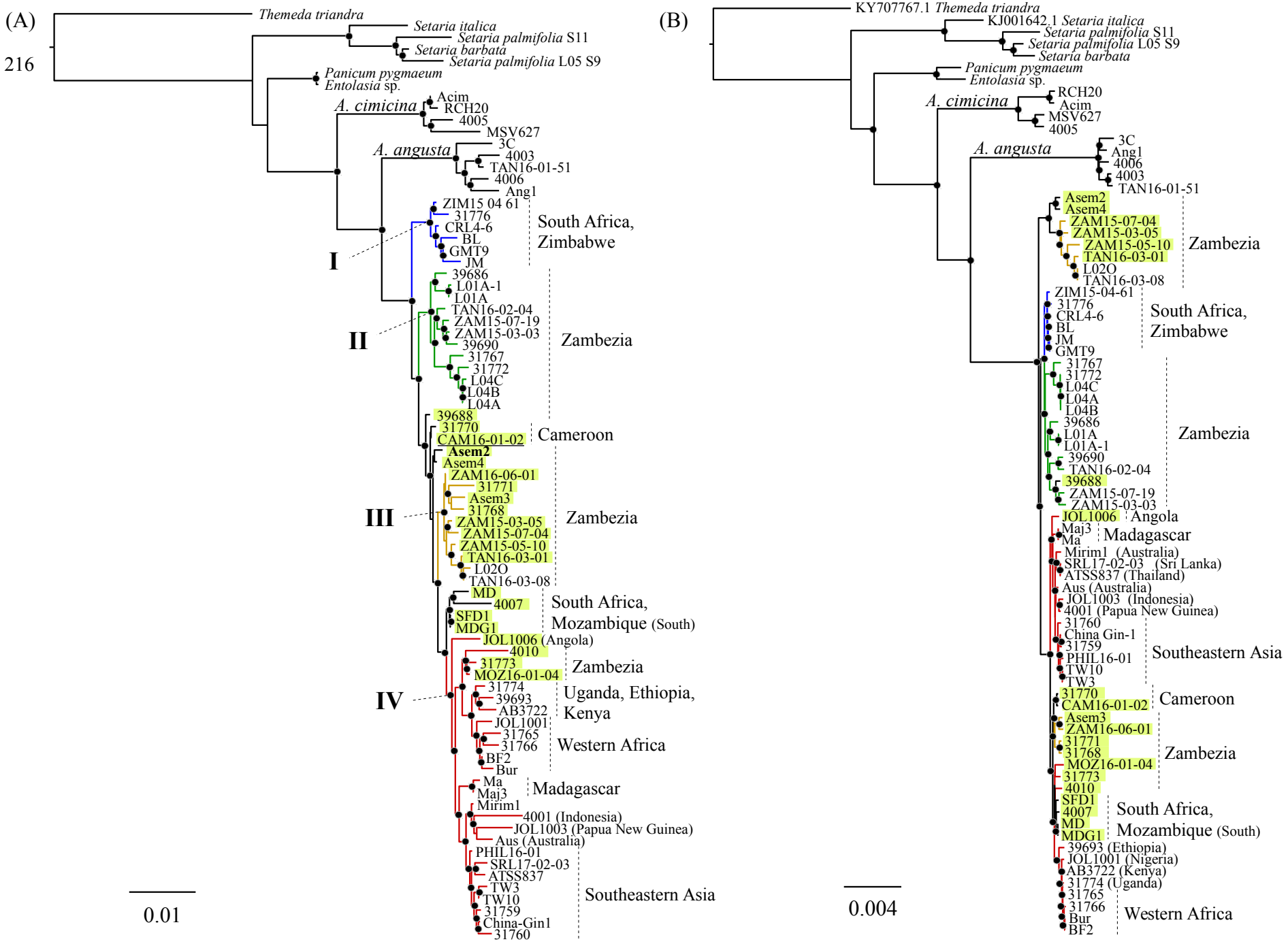

(A)
216
Themeda triandra
Setaria italica
Setaria palmifolia S11
Setaria barbata
Setaria palmifolia L05 S9
Panicum pygmaeum
Entolasia sp.
A. cimicina
Acim
RCH20
4005
MSV627
A. angusta
3C
4003
TAN16-01-51
4006
Ang1
ZIM15 04 61
31776
CRL4-6
BL
GMT9
JM
South Africa, Zimbabwe
I
39686
L01A-1
L01A
TAN16-02-04
ZAM15-07-19
ZAM15-03-03
39690
II
Zambezia
31767
31772
L04C
L04B
L04A
39688
31770
CAM16-01-02
Cameroon
Asem2
Asem4
ZAM16-06-01
31771
Asem3
31768
III
ZAM15-03-05
ZAM15-07-04
ZAM15-05-10
TAN16-03-01
L02O
TAN16-03-08
Zambezia
MD
4007
SFD1
MDG1
South Africa, Mozambique (South)
JOL1006 (Angola)
4010
31773
MOZ16-01-04
Zambezia
31774
39693
AB3722
IV
Uganda, Ethiopia, Kenya
JOL1001
31765
31766
BF2
Bur
Western Africa
Ma
Maj3
Madagascar
Mirim1
4001 (Indonesia)
JOL1003 (Papua New Guinea)
Aus (Australia)
PHIL16-01
SRL17-02-03
ATSS837
TW3
TW10
31759
China-Gin1
31760
Southeastern Asia
0.01
(B)
KY707767.1 Themeda triandra
KJ001642.1 Setaria italica
Setaria palmifolia S11
Setaria palmifolia L05 S9
Setaria barbata
Panicum pygmaeum
Entolasia sp.
A. cimicina
RCH20
Acim
MSV627
4005
A. angusta
3C
Ang1
4006
4003
TAN16-01-51
Asem2
Asem4
ZAM15-07-04
ZAM15-03-05
ZAM15-05-10
TAN16-03-01
L02O
TAN16-03-08
Zambezia
ZIM15-04-61
31776
CRL4-6
BL
JM
GMT9
South Africa, Zimbabwe
31767
31772
L04C
L04A
L04B
39686
L01A
L01A-1
39690
TAN16-02-04
39688
ZAM15-07-19
ZAM15-03-03
Zambezia
JOL1006
Angola
Maj3
Ma
Madagascar
Mirim1 (Australia)
SRL17-02-03 (Sri Lanka)
ATSS837 (Thailand)
Aus (Australia)
JOL1003 (Indonesia)
4001 (Papua New Guinea)
31760
China Gin-1
31759
PHIL16-01
TW10
TW3
Southeastern Asia
31770
CAM16-01-02
Cameroon
Asem3
ZAM16-06-01
31771
31768
MOZ16-01-04
Zambezia
31773
4010
SFD1
4007
MD
MDG1
South Africa, Mozambique (South)
39693 (Ethiopia)
JOL1001 (Nigeria)
AB3722 (Kenya)
31774 (Uganda)
31765
31766
Bur
BF2
Western Africa
0.004

Fig. 5.3. Maximum likelihood trees of *Alloteropsis* and other Panicoideae grasses based on (A) nuclear genome-wide markers and (B) full plastome sequences. Branches were coloured according to the nuclear clades of *A. semialata* (Chapter 4), with nuclear clades I, II, III and IV in blue, green, yellow and red, respectively. Accessions highlighted in green carry the laterally acquired PEPC gene (*ppc1P3_LGT:C*). Black circles on nodes are bootstrap support values > 50 (values ≤ 50 were omitted).

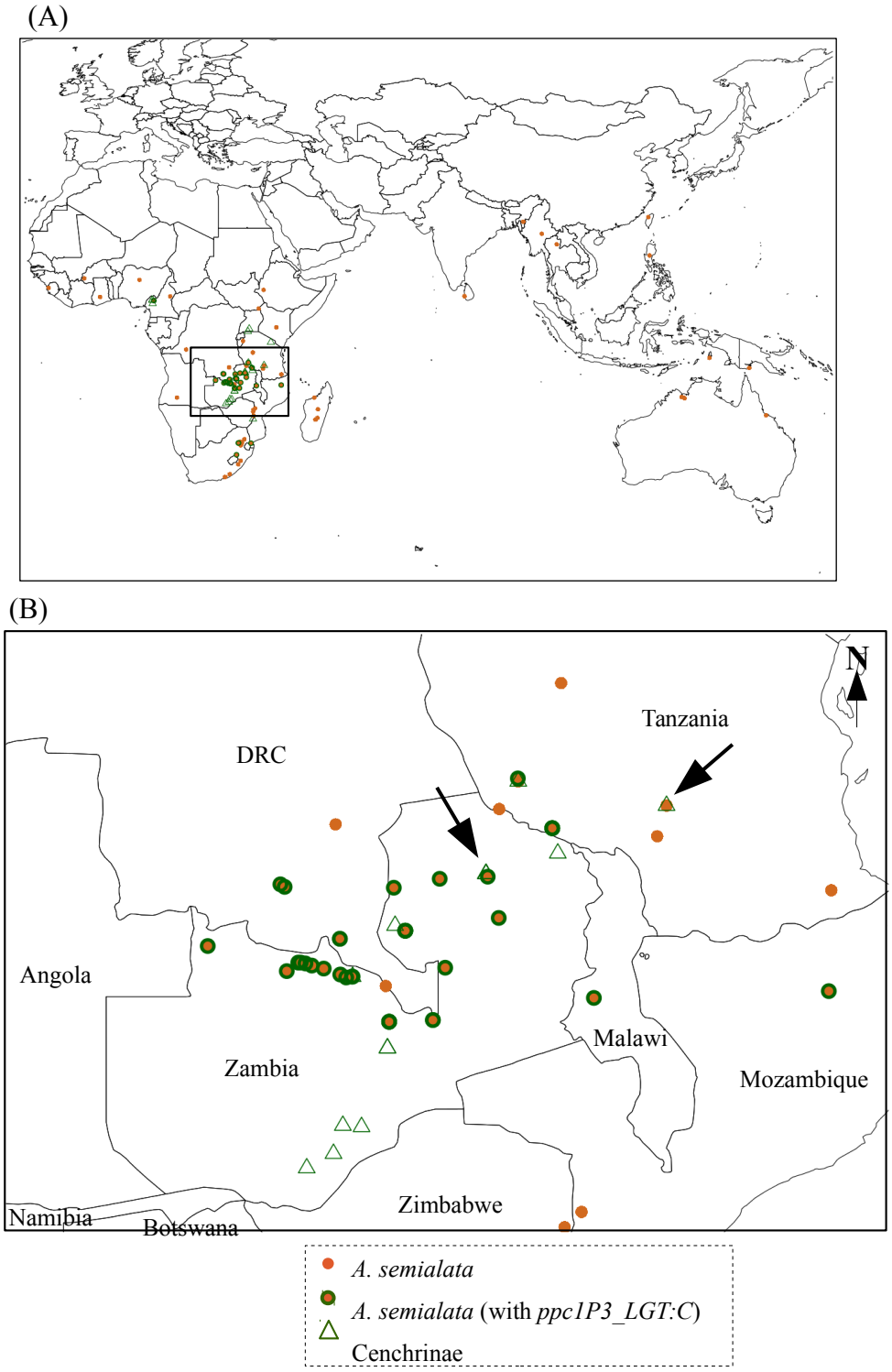


Fig. 5.4. Geographical distribution of *A. semialata* and Cenchrinae accessions analysed here. (A) World distribution; (B) detail of the Zambezian region of Africa. Arrows indicate Cenchrinae accessions having *ppc1P3_LGT:C* most closely related to *A. semialata* (i.e. SPC01 from Tanzania, and SPC16 from Zambia; see Fig. 5.2).

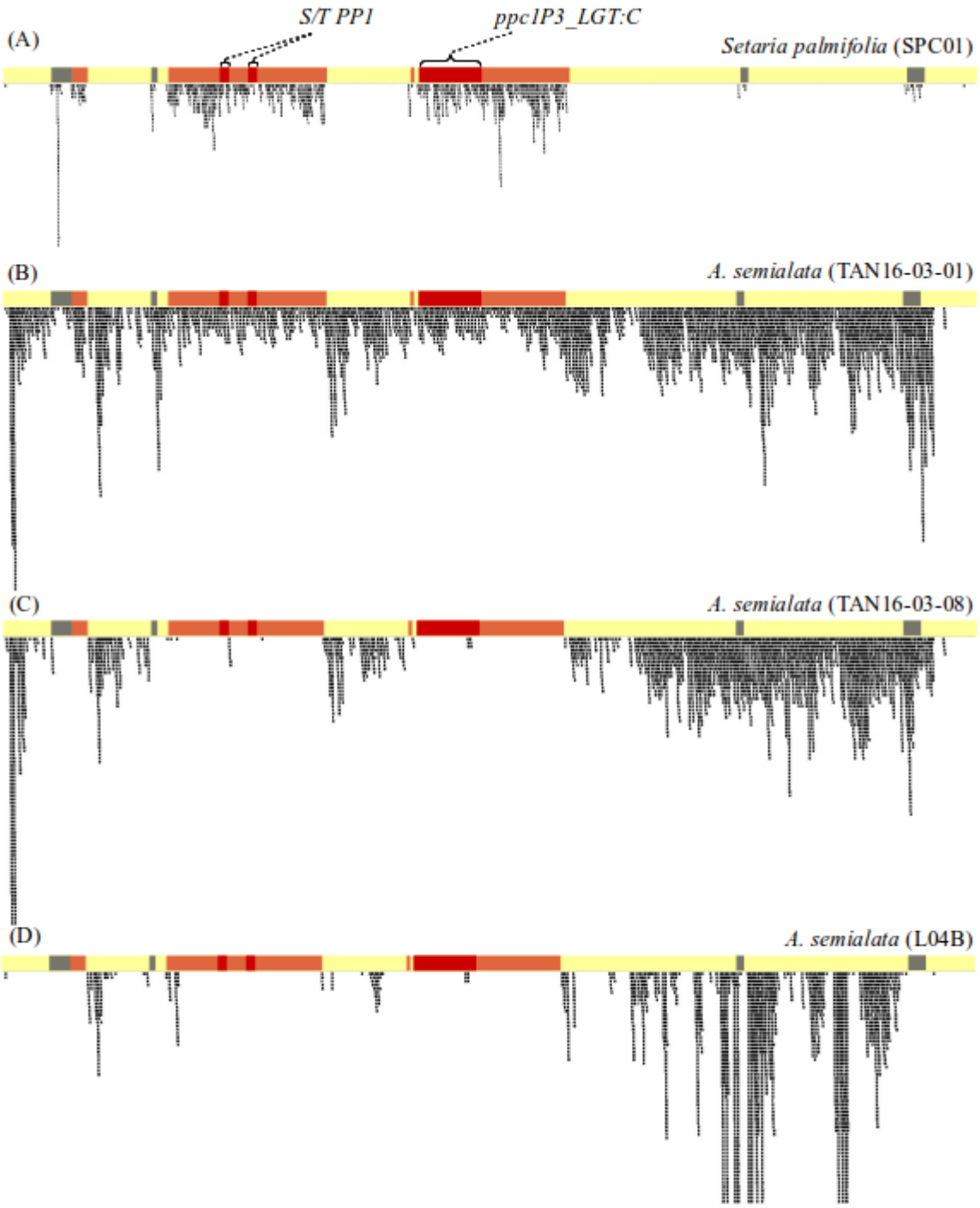
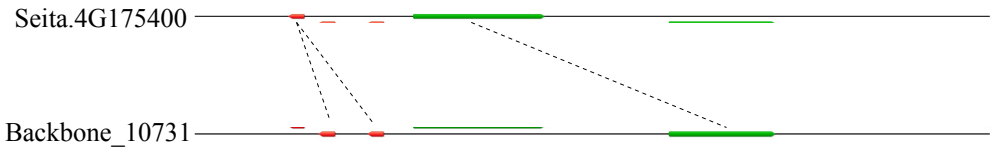


Fig. 5.5. Coverage plot illustrating the reference scaffold containing *ppc1P3_LGT:C*. (A) Member of the donor lineage, *S. palmifolia* (SPC01); (B) *A. semialata* accession carrying the LGT fragment (TAN16-03-01); (C) and (D) *A. semialata* accessions that do not carry the LGT fragments (TAN16-03-08 and L04B, respectively). Note intrapopulation variation in the presence/absence of the LGT fragments (B and C). Coloured bar on top of each coverage plot represents the reference genome scaffold that contains *ppc1P3_LGT:C*; yellow indicates segments that were vertically inherited, and orange and dark red indicate segments putatively laterally acquired from Cenchrinae; grey indicates segments with uncertain origin. Segments containing protein-coding genes are coloured in dark red (*ppc1P3_LGT:C* and tandem copies of *S/T PP1*).

(A)



(B)

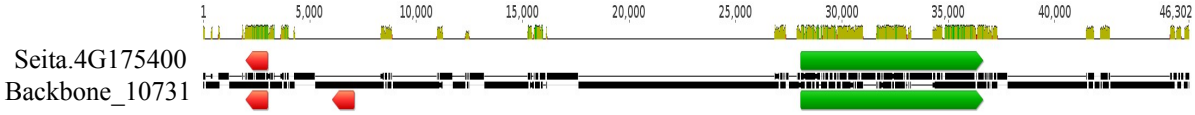


Fig. 5.6. Identity between genome scaffolds of *Setaria italica* (Seita.4G175400) and the reference genome of *A. semialata* (ZAM15-05-10; Backbone_10731). The region analysed is 50 kb long. (A) Schematic representation of the two scaffolds; (B) aligned scaffolds, with top bars indicate percentage of identity in the aligned regions: 100% (green), 30-100% (yellow) and < 30% (red). Protein-coding genes are annotated in green (*ppc1P3_LGT:C*) and red (*S/T PP1*).

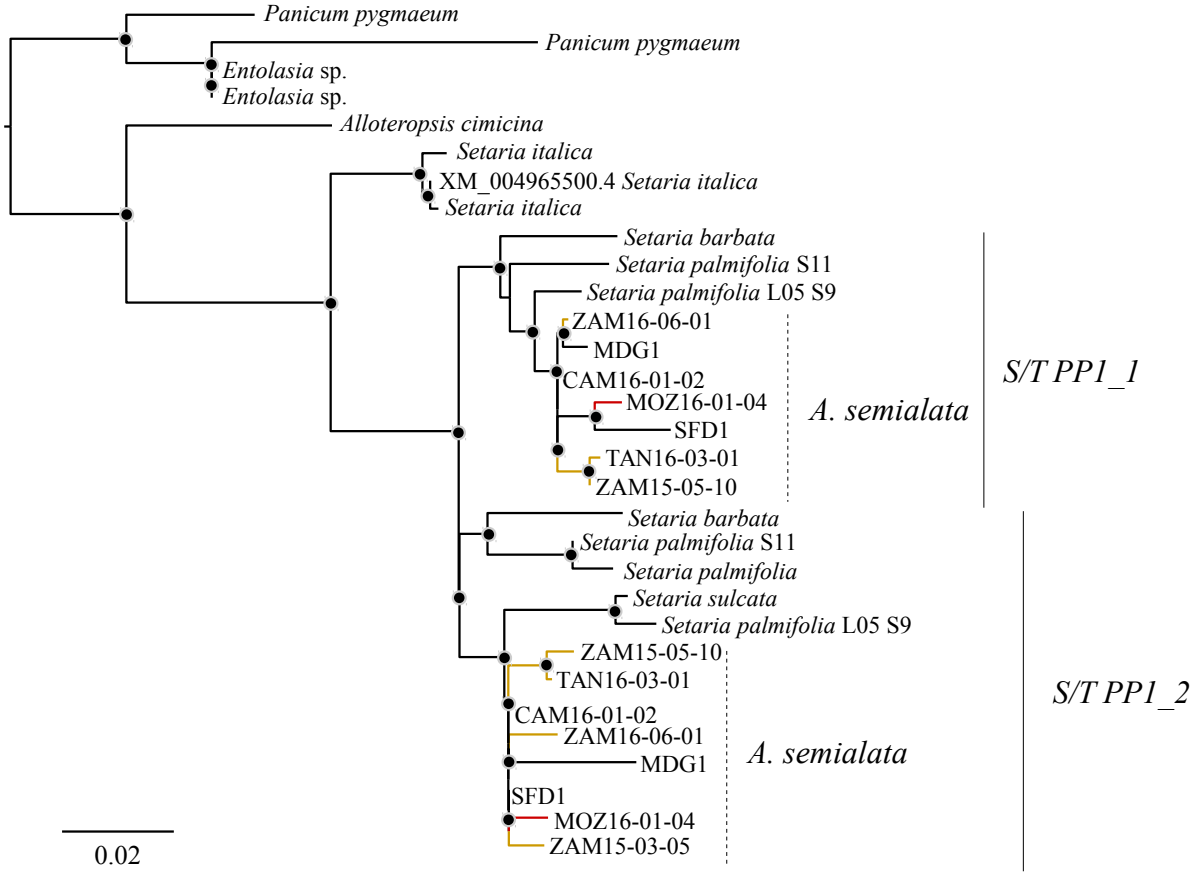


Fig. 5.7. Maximum likelihood tree of *S/T PP1*. Branches were coloured according to the nuclear clades of *A. semialata* (Chapter 4), with nuclear clades III and IV in yellow and red, respectively. Black circles on nodes are bootstrap support values > 50 (values ≤ 50 were omitted).

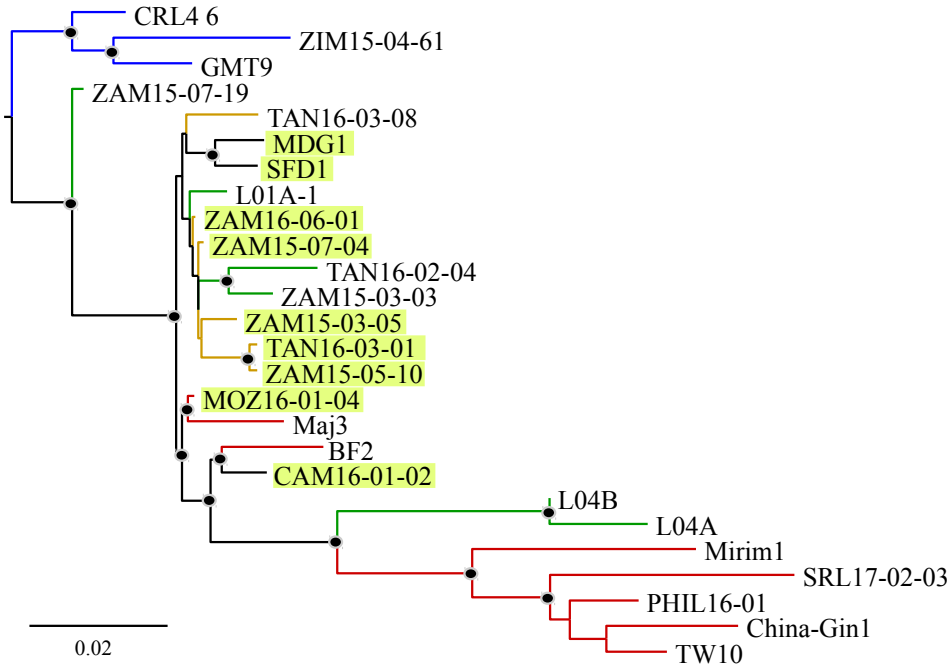


Fig. 5.8. Maximum likelihood tree based on the non-laterally acquired genomic segment separating *ppc1P3_LGT:C* and *S/T PP1* in the reference genome. Branches were coloured according to the nuclear clades of *A. semialata* (Chapter 4), with nuclear clades I, II, III and IV in blue, green, yellow and red, respectively. Accessions highlighted in yellow carry the laterally acquired PEPC gene (*ppc1P3_LGT:C*). Black circles on nodes are bootstrap support values > 50 (values ≤ 50 were omitted).

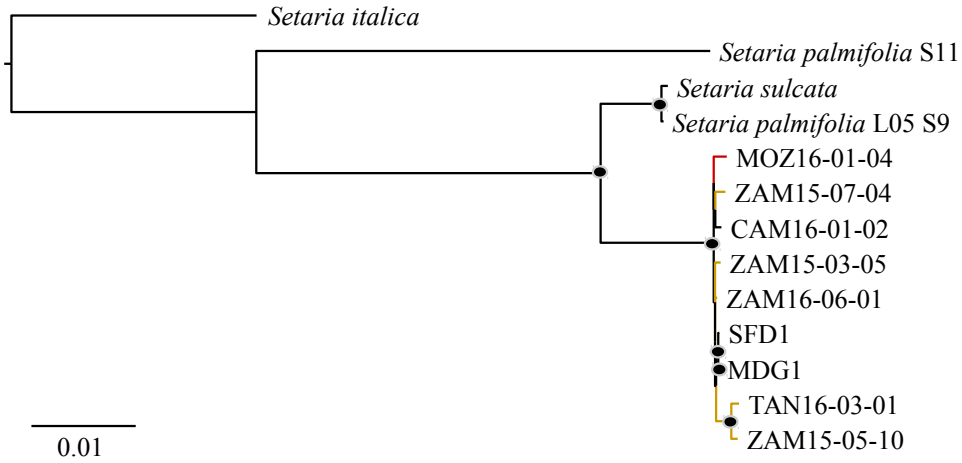


Fig. 5.9. Maximum likelihood tree of the genomic segment immediately 3' downstream to *ppc1P3_LGT:C*. Branches were coloured according to the nuclear clades of *A. semialata* (Chapter 4), with nuclear clades III and IV in yellow and red, respectively. Black circles on nodes are bootstrap support values > 50 (values ≤ 50 were omitted).

5.5. Discussion

5.5.1. The LGT event involving PEPC from Setaria *occurred in Central Africa*

Our phylogenetic analyses confirmed that the donor species is a member of the *Setaria palmifolia* complex within the Cenchrinae subtribe of grasses (Fig. 5.2; Christin et al. 2012a). The *S. palmifolia* complex comprises at least six perennial species distributed across Africa, Asia and America (Clayton 1979). We found an individual from Tanzania (SPC01) having a *ppc1P3_LGT:C* sequence highly similar to the *A. semialata* (Fig. 5.2). This individual grew a few meters apart from *A. semialata*. Another individual with very similar sequences (SPC23) came from a germplasm, originally from Puerto Rico. Members of the *S. palmifolia* complex have been imported in many places of the world as ornamentals, and later became invasive. It is consequently likely that these extra African accessions result from recent, human-mediated movements. The fact that sequences very similar to those detected within *Alloteropsis* were found in native populations of *Setaria* co-occurring in Tanzania with *A. semialata* shows that the LGT could have taken place in the Zambezian region.

Members of the *S. palmifolia* complex are often associated with shaded and moist habitats (Clayton 1979; Morrone et al. 2013), and are frequent in the Zambezian region. They include large understorey plants such as *S. megaphylla*, but members with thinner leaves were collected in wooded savannas. The morphological types do not match the genetic groups based on plastid markers, and different positions with respect to other Cenchrinae of the samples based on plastid and nuclear trees (Figs 5.1 and 5.2) support hybridization within the group as previously suggested (Clayton 1979). This makes the identification of a named species as the donor complicated, but members present in the wooded savannas of Tanzania and Zambia are excellent candidates.

5.5.2. The LGT fragment rapidly spread across A. semialata *populations*

Our analyses confirmed that *ppc1P3_LGT:C* is restricted to African populations of *A. semialata*, as previously suggested based on smaller sample size (Chapter 4). Most of the individuals belonging to the genomic group of the Zambezian region (particularly Tanzania, Zambia and DRC; nuclear clade III) possess the gene (Fig. 5.3). A second

group of individuals carrying the gene belong to the nuclear clade IV, which includes populations from South Africa, Mozambique, Malawi, Tanzania and Angola (Fig. 5.3). These are all C_4, and while the first group encompasses diploid individuals, the members of the second group for which genome sizes are available are polyploids (Table 5.S2). Finally, a third group of individuals carrying the gene are polyploids from Cameroon as well as individuals from DRC previously identified as admixed between nuclear clades II and III (Chapter 4). Since no other individuals from nuclear clades II or IV carry *ppc1P3_LGT:C*, we conclude that this gene was originally acquired either (1) before the split between nuclear clades III and IV, and subsequently lost in clade IV and one population from clade III, or (2) by members of the nuclear clade III (probably in the Zambezian region), after the split between African and Asia/Australian C_4 lineages of *A. semialata*. In the first scenario, the divergence of *ppc1P3_LGT:C* from members of clades III and IV would have happened more than two million years ago (Lundgren et al. 2015; Chapter 4), and the laterally acquired genes would therefore have accumulated mutations since the split of the two clades. The lack of variation in both laterally acquired genes on the fragment rules out the hypothesis (Figs 5.2 and 5.8), as shown previously based on molecular dating (Chapter 4). We therefore confirm that the gene was passed among established populations following its acquisition.

Our study reports for the first time the existence of individuals with and without the laterally acquired gene within the same populations. While the two populations in Zambia are composed of different photosynthetic types belonging to different genetic lineages (Fig. 5.3), the individuals with and without the gene from the Tanzanian populations belong to the same plastid and nuclear genomic groups and are very similar across their genome (Fig. 5.3). We therefore conclude that the observed polymorphism is recent. One possibility is that *ppc1P3_LGT:C* was recently suppressed from some individuals, but there is no reason why this would have led to the loss of the other LGT in the same fragment. It is therefore more likely that the gene reached the population recently, and that the polymorphism is transient. Transcriptome analyses indicate that *ppc1P3_LGT:C* is expressed in those individuals of the population that have it (Moreno-Villena et al. 2018). However, the same individuals express genes for the other laterally acquired PEPC (*ppc1P3_LGT:M*) at higher levels, which might limit the selective advantage of *ppc1P3_LGT:C*.

5.5.3. No evidence of selective sweep of the LGT fragment

Our genome analysis indicates that an additional gene was acquired with *ppc1P3_LGT:C* (Fig. 5.6). Altogether, the laterally acquired DNA sums up to 60kb. The two laterally acquired genes are separated by a region without similarity to *Setaria* species (Fig. 5.5). This segment was likely inserted after the LGT, during ongoing genome rearrangements. This process might have been accelerated by the insertion of TEs in regions surrounding both genes (Fig. 5.5). The immediate 5' upstream region of *ppc1P3_LGT:C* is not homologous between the reference genome and other *A. semialata,* nor *Setaria* accessions, suggesting a sequence insertion in the reference genome. However, homology search of a 3 kb segment using BLAST against NCBI public database showed the presence of transcription factor sequences, which might be associated with the regulation of *ppc1P3_LGT:C*. The regulatory mechanisms controlling the expression of laterally acquired genes are unknown and demand future studies.

The most likely scenario for the spread of *ppc1P3_LGT:C* as reported above would be selection-mediated introgression of the fragment. This would involve multiple rounds of hybridization (Chapter 4), followed by recombination and selection for the chromosomes bearing the *ppc1P3_LGT:C* gene. Such a typical case of selective sweep would move the flanking regions with the laterally acquired genes, because of linkage disequilibrium. The history of the LGT should therefore be shared with those of surrounding native DNA. However, the phylogenies based on the region between the two LGTs does not group the individuals with the LGT, and this region is more diverse than the LGTs (Fig. 5.8). We therefore conclude that, based on the available evidence, classical selection-mediated introgression is not responsible for the spread of the gene among populations of *A. semialata*. Reconciling the rapid spread of the gene as evidenced by its conservatism within *A. semialata* and the lack of genome-wide signatures requires further studies, but some hypotheses can be suggested. First, it is possible that the spread of the gene was followed by genomic rearrangements after hybridization. Admixed individuals are known in *A. semialata*, including polyploids (Chapter 4), but whether the set of chromosomes from different parents recombine is unknown. A scenario of translocation of the LGT among chromosomal regions in hybrids would explain the observed patterns of differentiation. Alternatively, it is possible that the LGT was incorporated independently in the genomes of different *A.*

semialata lineages. The vector for the transfers remain unknown, but extra chromosomal DNA fragments, such as eccDNA (Cohen et al. 2007; Lanciano et al. 2017) might have moved the genes from *Setaria* to *A. semialata*, as previously suggested for other LGTs involving this species (Dunning et al. in prep). These eccDNA fragments might then have spread to different populations, which would have integrated in different chromosomal locations. Testing these hypotheses requires follow-up investigations. In particular, FISH or genome sequencing of additional individuals would determine whether the LGT is in the same position in all *A. semialat*a accessions that have it.

5.6. Conclusions

Using a laterally acquired gene involved in a complex adaptive trait, we shed light on the dynamics of spread of a novel mutation in natural populations of the grass *A. semialata*. Members of the donor lineage overlap in their current distribution with populations of *A. semialata* in the Zambezian region of Africa. As previously suggested, the LGT is restricted to some African populations of *A. semialata*. However, its presence in distinct genomic lineages and its conservation across *A. semialata* populations suggests that it was acquired after the diversification of *A. semialata* and then spread across established populations. Analyses of a draft genome indicate that the gene was acquired together with another gene, making the LGT fragment at least 60kb long. Intriguingly, the separating region of native DNA is not more similar among individuals with the LGT than among close relatives with and without the LGT. This pattern is not compatible with a selective sweep involving rapid integration via hybridization. Instead, we suggest that the gene was transferred among established populations by means other than chromosomal recombination, using mechanisms that remain to be discovered.

5.7. Acknowledgements

We thank Oriane Hidalgo and Ilia Leitch for support with the genome size measurements, Luke T. Dunning and Jill K. Olofsson for providing support for the genome sequencing and analyses, Marjorie R. Lundgren for providing plant material, and Guillaume Besnard for providing genomic datasets.

5.8. Supporting Information

Table 5.S1. Genomic data information.

Data type	Species/Subtribe	Country	Accession	Latitude	Longitude	C[1]
PCR Sanger	*A. angusta*	Tanzania	TAN16-01-07	-7.94	31.68	
PCR Sanger	*A. semialata*	South Africa	ASM-1	-25.61	29.73	
PCR Sanger	*A. semialata*	South Africa	BLW-1	-29.71	29.96	
PCR Sanger	*A. semialata*	Cameroon	CAM16-01-09	5.93	10.62	X
PCR Sanger	*A. semialata*	South Africa	CRL-1	-25.74	30.24	
PCR Sanger	*A. semialata*	South Africa	EML-1	-26.29	30.00	
PCR Sanger	*A. semialata*	Madagascar	GB61.2014_1	-18.27	47.17	
PCR Sanger	*A. semialata*	South Africa	JMS-1	-33.32	26.44	
PCR Sanger	*A. semialata*	Australia	KRRd-1	-15.54	128.15	
PCR Sanger	*A. semialata*	South Africa	KSD-1	-30.51	29.43	
PCR Sanger	*A. semialata*	South Africa	KWT-1	-32.70	27.53	
PCR Sanger	*A. semialata*	Tanzania	L01-1	-5.63	32.69	
PCR Sanger	*A. semialata*	Tanzania	L02e (8/2-16)	-9.04	32.48	
PCR Sanger	*A. semialata*	Tanzania	L02f (23/3-16)	-9.04	32.48	X
PCR Sanger	*A. semialata*	Tanzania	L02g (23/3-16)	-9.04	32.48	
PCR Sanger	*A. semialata*	Tanzania	L02h (23/3-16)	-9.04	32.48	
PCR Sanger	*A. semialata*	Tanzania	L02i (23/3-16)	-9.04	32.48	
PCR Sanger	*A. semialata*	South Africa	LSU-2	-25.58	29.77	
PCR Sanger	*A. semialata*	South Africa	MDB-2	-25.74	29.50	
PCR Sanger	*A. semialata*	Australia	Mirim-2	-15.77	128.75	
PCR Sanger	*A. semialata*	Madagascar	MSV1935_1	-20.17	47.06	
PCR Sanger	*A. semialata*	Madagascar	MSV1937_1	-20.56	46.69	
PCR Sanger	*A. semialata*	Madagascar	MSV2081_1	-20.58	46.61	
PCR Sanger	*A. semialata*	South Africa	MTP-1	-25.62	30.26	
PCR Sanger	*A. semialata*	South Africa	PGR-1	-24.93	30.79	
PCR Sanger	*A. semialata*	South Africa	SFB-1	-25.48	29.79	
PCR Sanger	*A. semialata*	South Africa	SNR-1	-28.50	29.06	
PCR Sanger	*A. semialata*	Tanzania	TAN16-02-01	-7.94	31.68	
PCR Sanger	*A. semialata*	Tanzania	TAN16-02-01	-7.94	31.68	
PCR Sanger	*A. semialata*	Tanzania	TAN16-03-05	-9.04	32.48	
PCR Sanger	*A. semialata*	Tanzania	TAN16-03-09	-9.04	32.48	
PCR Sanger	*A. semialata*	Tanzania	TAN16-03-10	-9.04	32.48	
PCR Sanger	*A. semialata*	Tanzania	TAN16-03-144	-9.04	32.48	
PCR Sanger	*A. semialata*	Tanzania	TAN16-03-145	-9.04	32.48	X
PCR Sanger	*A. semialata*	Tanzania	TAN16-03-147	-9.04	32.48	
PCR Sanger	*A. semialata*	Tanzania	TAN16-03-148	-9.04	32.48	
PCR Sanger	*A. semialata*	Tanzania	TAN16-03-149	-9.04	32.48	
PCR Sanger	*A. semialata*	Tanzania	TAN16-03-150	-9.04	32.48	
PCR Sanger	*A. semialata*	Tanzania	TAN16-03-152	-9.04	32.48	X
PCR Sanger	*A. semialata*	Tanzania	TAN16-03-153	-9.04	32.48	
PCR Sanger	*A. semialata*	Tanzania	TAN16-03-154	-9.04	32.48	
PCR Sanger	*A. semialata*	Tanzania	TAN16-03-156	-9.04	32.48	

PCR Sanger	*A. semialata*	Tanzania	TAN16-03-162	-9.04	32.48	
PCR Sanger	*A. semialata*	Tanzania	TAN16-03-163	-9.04	32.48	X
PCR Sanger	*A. semialata*	Tanzania	TAN16-03-164	-9.04	32.48	
PCR Sanger	*A. semialata*	Tanzania	TAN16-04-92	-8.60	31.24	
PCR Sanger	*A. semialata*	Tanzania	TAN16-04B-117	-8.52	31.21	
PCR Sanger	*A. semialata*	Tanzania	TAN16-04C-129	-8.37	31.24	
PCR Sanger	*A. semialata*	Zambia	TS-A-01	-12.20	26.56	X
PCR Sanger	*A. semialata*	Zambia	TS-B-01	-12.20	26.50	X
PCR Sanger	*A. semialata*	Zambia	TS-C-01	-12.22	26.67	X
PCR Sanger	*A. semialata*	Zambia	TS-D-01	-12.27	26.82	X
PCR Sanger	*A. semialata*	Zambia	TS-E-01	-12.32	29.96	X
PCR Sanger	*A. semialata*	Zambia	TS-F-01	-12.34	27.10	X
PCR Sanger	*A. semialata*	Zambia	TS-I-01	-12.48	27.49	X
PCR Sanger	*A. semialata*	Zambia	TS-J-01	-12.55	27.63	X
PCR Sanger	*A. semialata*	Australia	XFRd-1	-15.83	128.79	
PCR Sanger	*A. semialata*	Zambia	ZAM15-01-10	-11.15	31.22	X
PCR Sanger	*A. semialata*	Zambia	ZAM15-02-01	-10.18	30.96	X
PCR Sanger	*A. semialata*	Zambia	ZAM15-03-08	-10.23	29.83	
PCR Sanger	*A. semialata*	Zambia	ZAM15-03-13	-10.23	29.83	
PCR Sanger	*A. semialata*	Zambia	ZAM15-03-69	-10.23	29.83	
PCR Sanger	*A. semialata*	Zambia	ZAM15-04-10	-10.44	28.75	X
PCR Sanger	*A. semialata*	Zambia	ZAM15-05-05	-11.45	29.02	X
PCR Sanger	*A. semialata*	Zambia	ZAM15-05-09	-11.45	29.02	X
PCR Sanger	*A. semialata*	Zambia	ZAM15-05a-01	-11.45	29.02	X
PCR Sanger	*A. semialata*	Zambia	ZAM15-05b-01	-11.45	29.02	X
PCR Sanger	*A. semialata*	Zambia	ZAM15-06-02	-13.55	29.67	X
PCR Sanger	*A. semialata*	Zambia	ZAM15-07-08	-11.81	24.37	
PCR Sanger	*A. semialata*	Zambia	ZAM15-08-09	-12.40	26.23	X
PCR Sanger	*A. semialata*	Zambia	ZAM15-09-01	-12.53	27.78	X
PCR Sanger	*A. semialata*	Zambia	ZAM15-09-07	-12.53	27.78	
PCR Sanger	*A. semialata*	Zambia	ZAM15-10-08	-13.59	28.64	X
PCR Sanger	*A. semialata*	Zambia	ZAM15-10-09	-13.59	28.64	X
PCR Sanger	*A. semialata*	Zambia	ZAM15-10-11	-13.59	28.64	X
PCR Sanger	*A. semialata*	Zambia	ZAM15-10-33	-13.59	28.64	X
PCR Sanger	*A. semialata*	Zambia	ZAM15-10-53	-13.59	28.64	X
PCR Sanger	*A. semialata*	Zambia	ZAM15-B-01	-12.53	27.76	X
PCR Sanger	*A. semialata*	Zambia	ZAM15-NP	-	-	X
PCR Sanger	*A. semialata*	Zimbabwe	ZIM15-02-09	-18.42	32.77	
PCR Sanger	*A. semialata*	Zimbabwe	ZIM15-03-03	-18.78	32.74	
PCR Sanger	*A. semialata*	Zimbabwe	ZIM15-04-08	-19.70	32.86	
PCR Sanger	Cenchrinae	Tanzania	SPC01[7*]	-8.51	35.17	-
PCR Sanger	Cenchrinae	Uganda	SPC02	-0.96	31.58	-
PCR Sanger	Cenchrinae	Tanzania	SPC03	-3.37	36.69	-
PCR Sanger	Cenchrinae	Uganda	SPC04	-0.36	31.87	-

PCR Sanger	Cenchrinae	Cameroon	SPC08	6.02	10.27	-
PCR Sanger	Cenchrinae	Cameroon	SPC09	5.25	10.37	-
PCR Sanger	Cenchrinae	Cameroon	SPC10	6.02	10.27	-
PCR Sanger	Cenchrinae	Zambia	SPC12	-10.11	30.92	-
PCR Sanger	Cenchrinae	Zambia	SPC16	-10.11	30.92	-
PCR Sanger	Cenchrinae	Zambia	SPC20	-12.53	27.78	-
PCR Sanger	Cenchrinae	Tanzania	SPC21	-7.94	31.68	-
PCR Sanger	Cenchrinae	Zimbabwe	SPC22	-20.42	32.71	-
PCR Sanger	Cenchrinae	-	SPC23[7**]	-	-	-
PCR Sanger	Cenchrinae	Zambia	SPC28	-11.33	28.78	-
PCR Sanger	Cenchrinae	Zambia	SPC29	-10.11	30.92	-
PCR Sanger	Cenchrinae	Zambia	SPC30	-10.11	30.92	-
PCR Sanger	Cenchrinae	Zambia	SPC31	-9.63	32.61	-
PCR Sanger	Cenchrinae	Zambia	SPC32	-14.21	28.60	-
PCR Sanger	Cenchrinae	Zambia	SPC33	-14.21	28.60	-
PCR Sanger	Cenchrinae	Zambia	SPC34	-16.03	27.55	-
PCR Sanger	Cenchrinae	Zambia	SPC36	-17.03	26.70	-
PCR Sanger	Cenchrinae	Zambia	SPC37	-16.69	27.33	-
PCR Sanger	Cenchrinae	Zambia	SPC38	-16.06	27.99	-
draft genome	*A. semialata*	Zambia	ZAM15-05-10	-11.45	29.02	X
resequencing[2]	*A. cimicina*	-	Acim	-	-	
resequencing[2]	*A. semialata*	Burkina Faso	BF2	10.85	-4.83	
resequencing[2]	*A. semialata*	Cameroon	CAM16-01-02	5.93	10.62	X
resequencing[2]	*A. semialata*	South Africa	CRL4-6	-25.74	30.24	
resequencing[2]	*A. semialata*	South Africa	GMT9	-33.32	26.53	
resequencing[2]	*A. semialata*	Tanzania	L01A	-5.63	32.69	
resequencing[2]	*A. semialata*	Tanzania	L04B	-8.51	35.17	
resequencing[4]	*A. semialata*	Tanzania	L04C	-8.51	35.17	
resequencing[2]	*A. semialata*	Madagascar	Maj3	-15.67	46.37	
resequencing[2]	*A. semialata*	South Africa	MDG1	-25.76	29.47	X
resequencing[2]	*A. semialata*	Australia	Mirim1	-15.77	128.75	
resequencing[2]	*A. semialata*	Philippines	PHIL16-01	15.95	121.01	
resequencing[2]	*A. semialata*	South Africa	SFD1	-28.39	29.04	X
resequencing[2]	*A. semialata*	Tanzania	TAN16-02-04	-7.94	31.68	
resequencing[2]	*A. semialata*	Tanzania	TAN16-03-08	-9.04	32.48	
resequencing[2]	*A. semialata*	Tanzania	TAN16-03-01	-9.04	32.48	X
resequencing[4]	*A. semialata*	Taiwan	TW10	24.47	120.72	
resequencing[2]	*A. semialata*	Zambia	ZAM15-03-03	-10.23	29.83	
resequencing[2]	*A. semialata*	Zambia	ZAM15-03-05	-10.23	29.83	X
resequencing[2]	*A. semialata*	Zimbabwe	ZIM15-04-61	-19.70	32.86	
resequencing	Cenchrinae	-	Setaria_italica (NCBI SRA)	-	-	-
resequencing[2]	*Entolasia* sp.	Australia	EM-AUS1	-26.57	150.55	
resequencing[2]	*Panicum pygmaeum*	-	PPyg2	-	-	
low-coverage	*A. angusta*	Tanzania	4003	-10.67	35.60	

low-coverage	*A. angusta*	Malawi	4006	-11.01	33.85	
low-coverage[3]	*A. angusta*	Uganda	3C	-0.36	31.87	
low-coverage[5]	*A. angusta*	DRC	Ang1	-4.04	21.76	
low-coverage	*A. angusta*	Tanzania	TAN16-01-51	-7.94	31.68	
low-coverage[3]	*A. cimicina*	Madagascar	RCH20	-18.77	46.87	
low-coverage	*A. paniculata*	Mozambique	4005	-19.71	34.79	
low-coverage[5]	*A. paniculata*	Madagascar	MSV627	-18.77	46.87	
low-coverage	*A. semialata*	Papua New Guinea	4001	-9.08	143.20	
low-coverage	*A. semialata*	Mozambique	4007	-25.81	32.27	x
low-coverage	*A. semialata*	Malawi	4010	-13.03	33.46	x
low-coverage	*A. semialata*	Myanmar	31759	20.78	97.03	
low-coverage	*A. semialata*	India	31760	23.48	92.89	
low-coverage	*A. semialata*	Sierra Leone	31765	8.77	-12.78	
low-coverage	*A. semialata*	Ghana	31766	6.73	-1.34	
low-coverage	*A. semialata*	DRC	31767	-8.95	27.38	
low-coverage[5]	*A. semialata*	DRC	31768	-11.64	27.48	x
low-coverage	*A. semialata*	Cameroon	31770	6.85	14.27	x
low-coverage	*A. semialata*	DRC	31771	-4.97	17.83	x
low-coverage	*A. semialata*	Tanzania	31772	-9.23	34.95	
low-coverage	*A. semialata*	Tanzania	31773	-10.50	39.05	x
low-coverage	*A. semialata*	Uganda	31774	4.12	33.98	
low-coverage	*A. semialata*	Mozambique	31776	-18.06	33.17	
low-coverage	*A. semialata*	Burundi	39686	-3.07	30.50	
low-coverage[5]	*A. semialata*	Tanzania	39688	-7.87	31.67	x
low-coverage	*A. semialata*	DRC	39690	-12.75	28.56	
low-coverage	*A. semialata*	Ethiopia	39693	8.29	35.05	
low-coverage[5]	*A. semialata*	Kenya	AB3722	-0.02	37.91	
low-coverage[5]	*A. semialata*	DRC	Asem2	-10.42	26.18	x
low-coverage[5]	*A. semialata*	DRC	Asem3	-11.64	27.48	x
low-coverage[5]	*A. semialata*	DRC	Asem4	-10.36	26.08	x
low-coverage[3]	*A. semialata*	Thailand	ATSS837	18.41	100.33	
low-coverage[3]	*A. semialata*	Australia	Aus	-19.62	146.96	
low-coverage[3]	*A. semialata*	South Africa	BL	-29.71	29.96	
low-coverage[3]	*A. semialata*	Burkina Faso	Bur	10.85	-4.83	
low-coverage	*A. semialata*	China	China	-	-	
low-coverage[3]	*A. semialata*	South Africa	JM	-33.32	26.44	
low-coverage	*A. semialata*	Nigeria	JOL1001	10.52	7.44	
low-coverage	*A. semialata*	Indonesia	JOL1003	-6.83	134.33	
low-coverage	*A. semialata*	Angola	JOL1006	-15.66	15.79	x
low-coverage[3]	*A. semialata*	Tanzania	L02O	-9.04	32.48	
low-coverage[3]	*A. semialata*	Tanzania	L04A	-8.51	35.17	
low-coverage[3]	*A. semialata*	Madagascar	Ma	-15.67	46.37	
low-coverage[3]	*A. semialata*	South Africa	MD	-25.76	29.47	x
low-coverage	*A. semialata*	Mozambique	MOZ16-01-04	-12.87	38.99	x

low-coverage	*A. semialata*	Sri Lanka	SRL17-02-03	6.9	79.8	
low-coverage[3]	*A. semialata*	Taiwan	TW3	24.47	120.72	
low-coverage	*A. semialata*	Zambia	ZAM15-07-04	-11.81	24.37	X
low-coverage	*A. semialata*	Zambia	ZAM15-07-19	-11.81	24.37	
low-coverage	*A. semialata*	Zambia	ZAM16-06-01	-13.55	29.67	X
low-coverage[6]	Cenchrinae	Tanzania	Setaria_palmifolia_L05_S9[7*]	-8.51	35.17	-
low-coverage[6]	Cenchrinae	-	Setaria_barbata[7**]	-	-	-
low-coverage[6]	Cenchrinae	Tanzania	Setaria_palmifolia_S11	-3.37	36.69	-

[1] Presence of the laterally acquired copy of PEPC from Cenchrinae (*ppc1P3_LGT:C*).

[2] Dataset retrieved from Dunning et al. (in prep).

[3] Dataset retrieved from Lundgren et al. (2015).

[4] Dataset retrieved from Chapter 3.

[5] Dataset retrieved from Chapter 4.

[6] Dataset retrieved from Park et al. (in prep).

[7*] and [7**] Same accessions that were used for both PCR/Sanger and whole genome, low coverage sequencing.

Table 5.S2. Genome size of *A.semialata* accessions.

Accession	Country	Nuclear clade[3]	Genome size 2C (Gb)
Aus1[1]	Australia	IV	2.20
BurkinaFaso3[1]	Burkina Faso	IV	1.95
CAM16-01-04	Cameroon	II-III	11.31
CAM16-01-06/2	Cameroon	II-III	11.86
Majunga1[1]	Madagascar	IV	2.05
PHIL16-09	Philippines	IV	1.81
EML11-200	South Africa	I	1.68
GMT3-1D	South Africa	I	1.83
JMS201[1]	South Africa	I	1.80
KWT3	South Africa	I	1.74
MDG8-3[1]	South Africa	IV	5.22
SRL17-02-03	Sri Lanka	IV	1.98
TW10[1]	Taiwan	IV	1.87
L01A[1]	Tanzania	II	2.19
L02[1]	Tanzania	III	2.01
L04A[1]	Tanzania	II	1.88
L04B-21	Tanzania	II	2.02
TAN16-02-05	Tanzania	II	1.97
TAN16-03-01	Tanzania	III	2.00
TAN16-03-06	Tanzania	III	1.96
ZAM15-03-02	Zambia	?	2.09
ZAM15-03-03	Zambia	II	2.07
ZAM15-03-04	Zambia	?	2.06
ZAM15-03-05A	Zambia	III	5.71
ZAM15-03-06	Zambia	?	2.35
ZAM15-03-08	Zambia	?	5.52
ZAM15-05-05	Zambia	?	1.94
ZAM15-05-10[2]	Zambia	III	2.18
ZAM15-07-01	Zambia	?	2.09
ZAM15-07-02	Zambia	?	2.00
ZAM15-07-04	Zambia	III	5.53
ZAM15-07-05	Zambia	?	2.07
ZAM15-07-06	Zambia	?	2.01
ZAM15-07-07	Zambia	?	1.98
ZAM15-07-07	Zambia	?	2.08
ZAM15-07-10	Zambia	?	2.12
ZAM15-07-11	Zambia	?	5.61
ZAM15-07-12	Zambia	?	2.10

ZAM15-07-13	Zambia	?	2.01
ZAM15-07-14	Zambia	?	2.08
ZAM15-07-15	Zambia	?	2.12
ZAM15-07-16	Zambia	?	5.16
ZAM15-07-19	Zambia	II	2.04
ZIM15-04-01	Zimbabwe	?	1.64

[1] Data retrieved from Lundgren et al. (2015).

[2] Sequenced genome.

[3] *A. semialata* lineages based on nuclear genome phylogeny (Chapter 4).

Chapter 6. General Discussion

6. General Discussion

In this work, I investigated the genomic changes associated with independent origins of the C_4 photosynthetic metabolism in grasses, which is a remarkable instance of repeated evolution of a complex trait in land plants. A comparative approach applied to a major C_4 lineage, the Andropogoneae grasses, suggests that the main components of an initial C_4 metabolism can evolve in a relatively short period, and that adaptive modifications of the trait are continuously accumulated during the diversification of a C_4 lineage (Chapter 2). Analyses of the genetic mechanisms underlying such phenotypic changes using a C_4 origin that is ~ 15 million years more recent, the grass *Alloteropsis*, uncovered a new role of gene duplication during C_4 evolution, via dosage effects creating rapid changes in expression levels of key C_4 genes (Chapter 3). In addition to the dosage effects of gene duplications, my analyses showed that the presence of laterally-acquired genes in some individuals coincides with the putative loss of ancient duplications (Chapter 3). My results are consistent with previous suggestions that these genes replaced the native copies, which were less suited to the C_4 function (Christin et al. 2012; Dunning et al. 2017). These findings provided the opportunity to investigate the population-level dynamics of novel adaptive mutations associated with the C_4 trait. Making use of vast genome-wide resources produced by our research group for *Alloteropsis*, I first tracked the dissemination of four laterally-acquired genes in *A. semialata*, and found that components of the C_4 trait can evolve in isolation and later be combined via gene flow (Chapter 4). Then, in order to gain insights into the genomic rearrangements and mode of spread of these adaptive mutations, I characterized the genomic fragment containing one of these laterally-acquired genes (Chapter 5). This analysis did not support the hypothesis of a rapid selective sweep of the entire fragment across populations via recurrent recombination, which urged for follow-up analyses to determine how the laterally-acquired fragment can be spread across genomic lineages. Overall, my investigations of genomic changes on different time scales suggest that (1) the components necessary for a rudimentary C_4 cycle are not many, and may be acquired in a relatively short period; (2) the complexity of the trait results from long periods of follow-up adaptation, and lineage-specific increments; (3) the genetic exchanges between divergent lineages can speed up the assembly and optimization of a C_4 metabolism.

6.1. A rudimentary C_4 cycle might be triggered by few genetic changes

C_4 photosynthesis is a biochemical cycle that emerges from multiple enzymes and membrane transporters, most of them with cell-specific expression patterns (Hatch 1987; Kanai and Edwards 1999). All these proteins were already present in the C_3 ancestors, performing general housekeeping functions (Leegood 2008; Aubry et al. 2011). Their co-option was, in many cases, followed by adjustments of their expression patterns and catalytic properties for a function in the mechanism of CO_2 fixation in leaves (Blasing et al. 2000; Tausta et al. 2002; Moreno-Villena et al. 2018). It might therefore be expected that massive changes in the protein sequences, regulatory mechanisms and cell metabolism underlie C_4 evolution. Such an impressive metabolic rewiring led to the hypothesis that numerous genetic changes are required for a C_4 cycle to be established (Sage 2004; Majeran et al. 2008; Gowik et al. 2011). However, the fact that C_4 evolved repeatedly and independently in multiple divergent lineages might suggest an easier trajectory, i.e. that a 'master' regulator would underlie the C_4 gene expression pattern. Whereas the first hypothesis of numerous parallel genetic changes has received some empirical support from differential gene expression analyses between C_3 and C_4 species (Bräutigam et al. 2011; Külahoglu et al. 2014; Wang et al. 2014), the hypothesis of a master regulator has not received any support. Here I present insights into this problem based on the macro and microevolutionary investigations conducted in this work.

Recent studies have shown that C_4 evolvability increases in lineages with particular enablers related to leaf anatomy (Sage 2001; Christin et al. 2013a), gene content (Monson 2003; Christin et al. 2009; Christin et al. 2013b, 2015; Dunning et al. 2017), and gene expression levels (Emms et al. 2016; Moreno-Villena et a. 2018). Such studies collectively suggest that the gap between non-C_4 and C_4 states is smaller than previously assumed since a considerable number of components attributed to the C_4 trait evolved before a C_4 cycle was established, and performed a different role in the whole plant physiology.

The hypothesis of 'exaptations' (i.e. adaptive features built by natural selection for a role that is different from their current one; Gould and Vrba 1982; Barve and Wagner 2013) facilitating C_4 evolution has an important ecological counterpart. Three major ecological factors are generally associated with C_4 evolution, namely (1) high

temperature, (2) low water availability, and (3) low atmospheric CO_2 (Ehleringer et al. 1997; Long 1999; Sage 2001; Osborne and Freckleton 2009). The drop in atmospheric CO_2 concentration during the Oligocene was particularly important for C_4 evolution (Pagani et al. 1999; Christin et al. 2008; Vicentini et al. 2008), as it led to significant levels of photorespiration in some terrestrial environments (Ehleringer et al. 1997). Warmth and aridity have affected the selective pressures on plants in some regions for long geological periods (Prentice et al. 1992; Williams et al. 2002; Woodward et al. 2004). This implies that the lineages that would give rise to C_4 plants accumulated adaptive traits associated with warm temperatures and low water availability during millions of years (Osborne and Sack 2012). Considering that (1) the pervasiveness of such factors increases the likelihood of some lineages convergently evolving the same adaptations, and (2) a subset of these adaptations might be recruited for the C_4 trait, the exaptation hypothesis, when analysed in an ecological framework, provides an explanation for the apparent paradox of repeatedly evolving a complex trait. However, it still does not establish the number and order of changes that lead to the establishment of a C_4 cycle, nor their time-scale.

In Chapter 2, I identified Jansenelleae as a C_3 lineage sister to a major C_4 group, the Andropogoneae, and showed that this C_3 group does not have clear C_4 enablers. Their leaf anatomy is typical of C_3 grasses in terms of C_4-related characters, with two layers of bundle-sheath cells (BSC), large interveinal distances, and no evidence of Rubisco activity in BSC. In addition, their enzymes carry no signatures of C_4-adaptive changes. If one assumes that these modifications were not secondarily lost during the diversification of Jansenelleae, they then must have evolved in Andropogoneae only after the split between the two groups. Tests for adaptive enzyme evolution and analyses of the rate of protein evolution show that massive C_4-adaptive changes indeed happened after the divergence between the two groups (Chapter 2). In addition, the loss of one BSC layer also postdates this divergence, and probably happened concomitantly with the enzymatic changes during the 3-4 millions years period before the first split within Andropogoneae. This implies that the major C_4-related genomic changes can happen in a relatively short period of time.

One possible explanation for the relative ease of C_4 evolution in some groups might be that strong epistasis controls the components of C_4 biochemistry. In this case, specific changes on one or a few loci (e.g. leading to increased and/or cell-specific gene expression) would interact with other loci and result in phenotypes that are different

from the simple independent effect of each locus (Kroymann and Mitchell-Olds 2005; Phillips 2008; Shao et al. 2008). A possible illustration of that would be a sudden increase in the expression of the first carboxylase, PEPC, which would lead to a metabolic imbalance due to the accumulation of C_4 acids in the cell. This could trigger plastic responses of constitutive metabolic pathways that would operate specifically to restore the stoichiometric balance in the cell. A similar hypothesis was first proposed by Mallmann et al. (2014), which used biochemical modelling to show that imbalances in the N metabolism due to a C_2 metabolism could trigger the activity of enzymes that are part of the C_4 cycle. This explanation would ease the transition from a C_2 to a full C_4 metabolism; however, there is no current evidence that the establishment of a C_2 cycle is a necessary intermediate step between C_3 and C_4 plants. Therefore it is possible that mutations of large effect might promote a metabolic rewiring in both a C_3 and a C_2 context. Examples of such mutations would include (1) increased expression levels due to dosage effects of gene duplication (Mouchès et al. 1986; Chen et al. 2008; Cook et al. 2012; Chapter 3); and (2) introgression of an allele/gene with both expression patterns and catalytic properties that can be drastically different from that of its homologs (Christin et al. 2012; Dunning et al. 2017). I indeed showed here that the acquisition by non-C_4 individuals of laterally-acquired copies of two core C_4 genes (*ppc* and *pck*) via introgression might have triggered a weak C_4 cycle in some lineages of *A. semialata* (Chapter 4). In cases where these novel biochemical pathways related to CO_2 fixation lead to increased fitness (e.g. gain in biomass), these genetic changes could rapidly be fixed (Heckmann 2013, 2016).

The hypothesis developed above suggests that, with the ecological drivers in place and the possibility of co-opting characters for the C_4 function, a few mutations of large effect on key genes could suffice to trigger an initial C_4 cycle. Multiple recent studies have indeed provided evidence that major phenotypes associated with C_4 photosynthesis are controlled by a single or a few loci. Using recently diverged C_3, C_3-C_4 intermediates and C_4 populations of the grass *A. semialata*, Lundgren et al. (in prep) showed that a single leaf anatomical change – prolonged differentiation of minor veins – is responsible for multiple aspects of the C_4 anatomical phenotype. Genetic studies have furthermore showed that the constitutive expression of a single gene (*GOLDEN2-LIKE*) in the C_3 plant rice induces multiple aspects of the intracellular C_4 anatomy (Wang et al. 2017), while elevated biosynthesis of auxin increases vein density (Huang et al. 2017). Finally, Reyna-Llorens et al (2018) showed that regulatory motifs present

in the coding sequences are common to a number of core C_4 genes, and conserved across land plants, which reduces the regulatory gap necessary for C_4 gene expression. Current efforts to engineer the C_4 trait in C_3 plants involve constantly more sophisticated experimental designs to identify the key components that can initiate a rudimentary C_4 cycle (Kajala et al. 2011; Covshoff and Hibberd 2012; von Caemmerer et al. 2012; Huang et al. 2016). However, most of these efforts are concentrated in establishing the necessary genetic elements when considering C_3 species that are distantly related to C_4 lineages, such as rice. Rice belongs to a grass lineage (BEP) that diverged ~ 50-60 million years ago from the lineage that gave rise to all C_4 origins in grasses (i.e. PACMAD; Christin et al. 2014), so it probably lacks all or most of the C_4 enablers (Christin and Osborne 2013). This implies a larger gap to the C_4 phenotype, so that additional components might have to be introduced in order to achieve a full C_4 cycle. Future studies aiming at identifying key mutations for C_4 evolution should rather focus on genetic manipulations of C_3 plants closely related to C_4 lineages.

6.2. Lineage-specific features and the complexity of the C_4 phenotype

Complex biological features evolve gradually via successive modifications from a basic structure and/or metabolic function. Camera eyes evolved from generic photoreceptor structures, with subsequent modifications in descendant lineages (Arendt 2003; Gehring 2004; Lamb et al. 2007). The evolution of insect wings followed a similar pattern, i.e. an initial structure probably co-opted for gliding was subsequent modified for different purposes, for example the elytra in Coleoptera (Kukalová-Peck and Lawrence 1993; Tomoyasu et al. 2009). Instances such as these suggest that complexity gradually emerges from successive increments of adaptive components on top of an already functional, adapted system. As any complex trait, the emergence of the complex C_4 syndrome composed of numerous biochemical and structural components can also be explained in such terms.

C_4 photosynthesis is often presented as a categorical and uniform trait, in the sense that C_3, C_4 and CAM (Crassulacean Acid Metabolism) are alternative categories of photosynthetic metabolism. However, it is clear that C_4 and CAM are not alternatives to the C_3 cycle, but accessory functions built upon the C_3 cycle to increase CO_2 concentration at the site of Rubisco (Kellogg 1999). It is also clear that the C_4 trait is not

a uniform set of features, but a very heterogeneous trait with several anatomical and biochemical variants around a single theme (Hattersley 1984, Sinha and Kellogg 1996; Kellogg 1999; Furbank 2011; Lundgren et al. 2014, 2015; Dunning et al. 2017; Lundgren et al. in prep; Reeves et al. 2018; Chapter 2). Such heterogeneity is related to the highly convergent nature of the C_4 trait, which independently evolved in more than 62 lineages of angiosperms (Sage et al. 2011). The fact that each lineage has its unique evolutionary history, before and after C_4 evolved, contributes to the observed diversity across C_4 lineages.

Most existing models of C_4 evolution hypothesize that components of the C_4 metabolism were acquired gradually following a single, linear trajectory that culminates into the establishment of a full C_4 cycle (Sage 2004; Gowik and Westhoff 2011; Sage et al. 2012, 2018). Such models suggest that a C_2 metabolism is a necessary step for a C_4 biochemistry to evolve. However, this hypothesis is mostly based on phylogenetic and phenotpyic analysis of a single eudicot group, the genus *Flaveria*, and it does not account for the heterogeneity that exists in the trait across eudicot and monocot C_4 lineages. Here I propose a generalization of this model based on the current knowledge on C_4 evolution (Fig. 6.1). A relative time-scale for the different events during the evolutionary trajectory is suggested. In addition, once C_4 evolution from a non-C_2 background cannot be excluded, an evolutionary bypass of the C_2 metabolism as an alternative trajectory is included.

My new model represents a phenotypic continuum, along which evolution can proceed potentially in several directions. This mirrors the approach of Brautigam and Gowik (2016), although the model that emerges from my work differs from previous efforts in important aspects. First, it is crucial to acknowledge that the independent C_4 origins are unlikely to have all followed the exact same trajectory, and the existence of a variety of paths across the adaptive landscape might have contributed to the high number of origins of the trait. The variety of trajectories between C_3 and C_4 states was already emphasized by Williams et al. (2013) and their phenotypic models, and to some extent recognized by Heckmann et al. (2013), who acknowledged that C_4 must not necessarily proceed via the C_2 cycle. However, almost all previous efforts to infer evolutionary trajectories considered the C_4 phenotype as a single stage, which represents the end point of evolutionary trajectories. This led to a lack of interest in the changes that happen once the C_4 phenotype emerged, and also the wide reliance on only one species per C_4 lineage. The importance to capture the diversity within each C_4 lineages

was well established among C_4 ecologists and physiologists (Edwards and Still 2008; Edwards and Smith 2010; Osborne and Freckleton 2009; Atkinson et al. 2016), but the *Alloteropsis* studies conducted in this thesis (Chapters 3, 4 and 5) and by other members of our research group (Dunning et al. 2017) were the first to consider evolutionary trajectories where a multitude a C_4 'end points' exist. Applying the same approach to the large Andropogoneae C_4 group clearly showed that most C_4 characteristics have been acquired after the C_4 physiology emerged (Chapter 2; Besnard et al. 2018), so that the complexity of the C_4 trait as observed in some extant species results from millions of years of adaptation of the existing C_4 machinery.

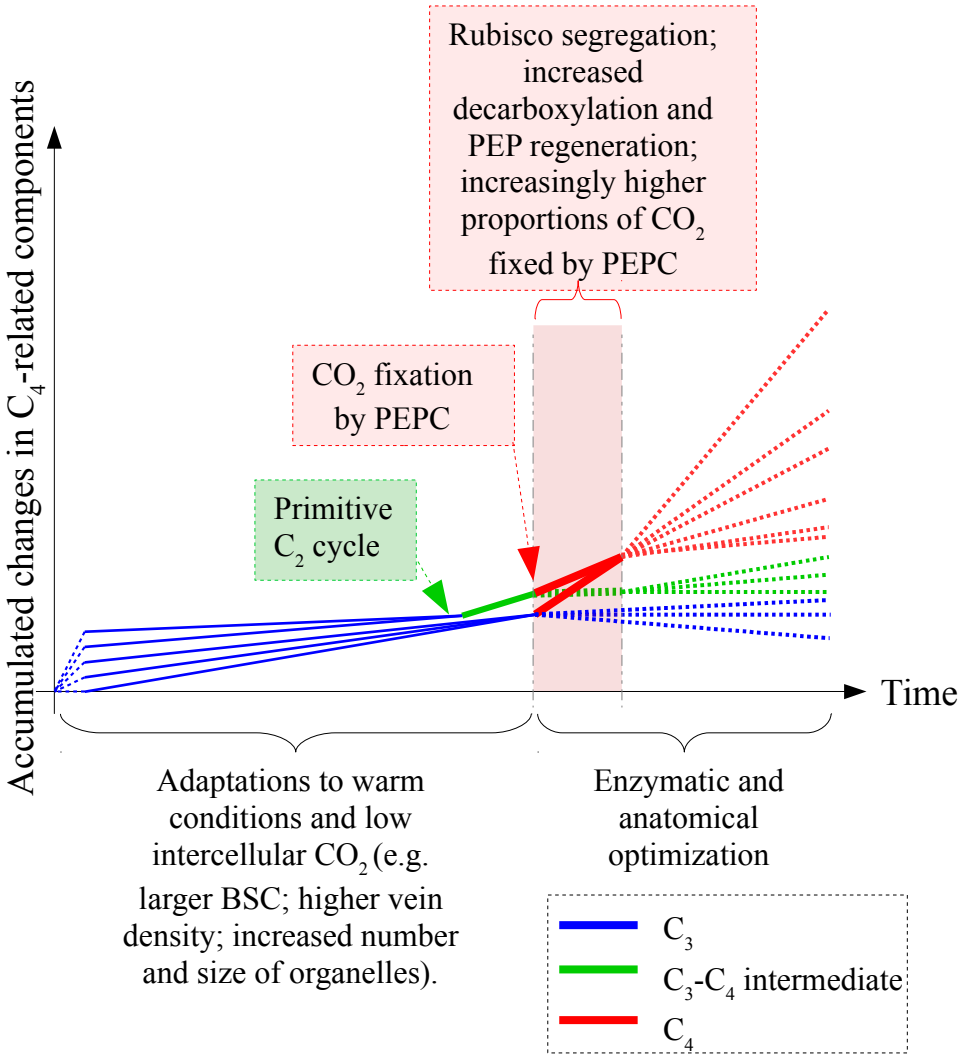


Fig. 6.1. A simplified model for the evolution of C_4 photosynthesis. A C_4 phenotype can evolve from multiple trajectories. Long periods of accumulation of C_4-related characters may precede C_4 evolution. A C_4 cycle might evolve from both C_2 and C_3 biochemical backgrounds. Multiple C_4 phenotypes originate from lineage-specific optimizations of leaf anatomical and biochemical traits.

6.3. Gene flow between divergent lineages as a source of novel adaptive mutations

Isolation promotes genetic divergence by allowing distinct gene pools to build up. Divergent lineages accumulate genetic variants, which can be neutral and fixed through drift or have been selected in a process of local adaptation (e.g. Dionne et al. 2008; Fournier-Level et al. 2011; Papadopulos et al. 2014). Secondary contacts between isolated gene pools can be a source of advantageous mutations (Whitney et al. 2010; Hedrick 2013; Stern 2013; Pardo-Diaz et al. 2012; Huerta-Sánchez et al. 2014; Kreiner et al. 2018). Here I provide evidence for an instance of introgression between isolated populations providing adaptive genes related to the C_4 trait (Chapter 4), as well as further evidence for the importance of gene flow between even more distantly related grass lineages, in the form of LGTs, for the origin and optimization of C_4 photosynthesis (Chapters 2, 3 and 4).

Components of the C_4 trait vary considerably within *A. semialata*, not only between photosynthetic types, but also within the C_4 group. Different genes have been co-opted for the C_4 function in different C_4 subgroups of *A. semialata* (Dunning et al. 2017; Dunning et al. in prep), and this is accompanied by large variation in C_4 gene expression levels (Dunning et al. in prep) and anatomical traits (Lundgren et al. 2016, in prep). Such differences indicate the accumulation of variation in C_4-related traits in divergent lineages of *A. semialata*, providing the basis for adaptive evolution via secondary contact between divergent lineages. In Chapter 4, we indeed provided evidence of gene flow between natural populations of divergent lineages of *A. semialata*, which might have contributed to the origin or strengthening of a weak C_4 cycle in the individuals with admixed genetic background. Coupling further genome-wide analysis with phenotype characterizations will provide an unique opportunity to track the within-species dynamics of genes associated with the C_4 metabolism. The current availability of two genomes of *A. semialata* (Dunning et al. in prep; Chapter 5) will facilitate such efforts.

While my work has highlighted the importance of hybridization for C_4 adaptation, this evidence came solely from the *A. semialata* species complex. The LGTs reported here similarly concerned *A. semialata*, as in previous studies (Christin et al. 2012, Dunning et al. 2017). This raises the question of whether some properties of *A. semialata* make genetic exchanges following secondary contacts especially important

for adaptation in this species. The genetic structure of *A. semialata* suggests highly fragmented populations (Chapter 4; Olofsson et al. In prep), which might have facilitated local adaptation. However, it is likely that secondary genetic exchanges facilitated C_4 adaptation in other groups as well. It has recently been suggested that C_3-C_4 intermediates might be of hybrid origin although there is no supporting evidence for this hypothesis (Kadereit et al. 2017). There is however evidence of transfer of C_4 genes among distinct species. First, genes for PCK, which were transferred between a member of *Cenchrus* and *Alloteropsis* (Christin et al. 2012; Chapter 4), were also apparently laterally-acquired by members of the grass genera *Echinochloa* and *Cymbopogon*, which both use it for their C_4 pathway (Moreno-Villena et al. 2018; Dunning et al. In prep). In addition, discrepancies between gene and species trees have suggested that genes for C_4 PEPC were somehow passed among members of the sedge genus *Eleocharis* (Besnard et al. 2009) and among members of the grass genus *Neurachne* (Christin et al. 2012). These cases were detected incidentally, in the absence of dedicated effort. It is therefore likely that targeted searches coupled with accumulating genomic resources will show that selection following genetic exchanges played an important role in C_4 evolution in many groups, supporting the overall importance of gene flow in the origins of adaptive innovations.

6.4. Applications of whole genome sequencing at low coverage

Although whole-genome sequencing costs have drastically decreased during the last 10 years (van Dijk et al. 2014), conducting comparative genomic studies can still be costly due to the requirement of large sample sizes. One alternative to reduce these costs is to pool many samples in a single sequencing run (Straub et al. 2012; Buerkle and Gompert 2013; Dodsworth 2015). Whole-genome sequencing at low coverage has been widely used for phylogenetic analyses (Besnard et al. 2013; Lundgren et al. 2015; Silva et al. 2016; Arthan et al. 2017; Hackel et al. 2018). Such studies mainly focus on the organellar portion of the genome, which are typically sequenced at higher depth because they are present in multiple copies within each cell. They are therefore easy to assemble and analyse and can provide useful phylogenetic markers. However, the reads obtained from the nuclear genome can still be used for many purposes (e.g. Besnard et al. 2014, 2018). These reads are generated as a by-product of organelle sequencing, but are

generally ignored, so that these resources exist for an increasing number of species, but remain largely unexplored.

The work presented in this thesis largely made use of available low coverage datasets, which in most cases were initially generated for phylogenetic analyses using plastid sequences (e.g. Lundgren et al. 2015; Arthan et al. 2017; Hackel et al. 2018). I developed novel approaches for phylogenomic analyses of the nuclear genome (Chapter 2), the analysis of gene copy numbers (Chapter 3), and phylogeographic analysis, based on the population genetic framework developed by Dr. Jill Olofsson (Chapter 4). I further collaborated on the analysis of specific C_4 genes using similar low-coverage sequencing datasets (Besnard et al. 2018). Previous studies have highlighted some of the advantages of using low-coverage sequencing, including its lower costs and the possibility to work on poorly preserved samples stored in museum collections. This can include rare species located in remote places (Chapter 2), individuals of some species from regions where field work is difficult due to unstable political situations (Chapter 4), or even species that are now extinct (Zedane et al. 2016). While the usefulness of low-coverage sequencing for phylogenetic analyses of organellar genomes, ribosomal DNA or specific nuclear marker was known, my work has shown that functional genomics is possible when combining such dataset with whole genomes or transcriptomes available for a few samples (Chapters 2, 3 and 4).

Overall my thesis contributed with novel approaches to extract robust information from the nuclear genome portion of low-coverage sequencing datasets. Because similar datasets are now routinely generated by many research groups, the number of species that can be included in such studies is becoming significant. There are now hundreds of grass plastomes in public databases. The corresponding sequencing reads are not generally released, but coordination among research groups allowed assembling a large species pool in Chapter 2. In the future, applying my approaches to the continuously generated low-coverage sequence data will allow conducting genomic studies on large scales. This will remove the problem of C_3/C_4 comparisons relying on a few species highlighted above. In addition, this will allow generating large-scale nuclear phylogenies for multiple groups, enabling explicit quantification of gene flow among species (e.g. Chapter 4). While such research program will undoubtedly shed new light into the origins of C_4 photosynthesis, my methods can be applied to any group and any trait of interest, therefore contributed to the advent of broad comparative genomic analyses of the origin of novel adaptations.

Conclusions

In this dissertation, I used genomic analyses coupled with phylogenetic methods to shed new light on the evolutionary origins of a complex physiological trait, the C_4 photosynthetic metabolism. Four hypotheses concerning the macro- and microevolutionary aspects of C_4 evolution were tested and provided novel insights into the subject. First, I showed that the evolution of some components of the C_4 trait can evolve via bursts of genetic changes concentrated in relatively short periods during the diversification of a C_4 lineage, as opposed to the view of gradual change over time. I also showed that such bursts of change were followed by continued adaptive evolution of some C_4 enzymes and further anatomical specializations. This finding indicates that the establishment of a C_4 cycle might not be the end of a single-route optimization trajectory, but the opening of numerous novel adaptive trajectories, in terms of both physiology (trait machinery) and ecology (niche specialization). Several genetic mechanisms and population-level processes that take place in shorter time scales may underlie such macroevolutionary patterns. Here I showed that gene duplications via dosage effects might provide an evolutionary shortcut to achieve the gene expression patterns and enzyme kinetics required for the C_4 pathway, as an alternative to the gradual accumulation of changes in regulatory and coding regions. Finally, analysis of the intraspecific dynamics of novel C_4-adaptive mutations indicated that components of the C_4 trait can evolve in isolated gene pools and later be combined via gene flow between individuals belonging to divergent genetic lineages. Further genomic analyses of one of such C_4-adaptive mutations did not confirm the hypothesis of its rapid sweep across populations, which calls for follow-up studies. In summary, the findings that I present in this dissertation reveal mechanisms and patterns that contribute to reduce the gaps that produce the apparent conundrum of evolving complex physiological traits. Numerous gaps in our knowledge still remain, and these are mainly related to the microevolutionary aspects of C_4 evolution. For example, the origins of tissue-specific expression patterns of C_4-related genes remain largely unknown. Successful crosses between closely related C_4 and non-C_4 individuals are key in this context, so that the genetic architecture of the C_4 trait can be further explored. Future studies should also test whether C_4 expression patterns can be achieved plastically via the induction of one or a few changes (e.g. PEPC and PCK). The grass *Alloteropsis semialata* is probably the

best candidate for such efforts, as it includes recently diverged C_4 and non-C_4 populations, and the work presented here has demonstrated the practicality of the system for both comparative genomics and analyses of gene flow and population dynamics.

References

Adams KL, Wendel JF. 2005. Polyploidy and genome evolution in plants. Curr. Opin. Plant Biol. 8:135–141.

Aird D, Ross MG, Chen W–S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biology 12:R18.

Akyildiz M, Gowik U, Engelmann S, Koczor M, Streubel M, Westhoff P. 2007. Evolution and function of a cis–regulatory module for mesophyll–specific gene expression in the C_4 dicot *Flaveria trinervia*. The Plant Cell 19:3391–3402.

Alkan C, Kidd JM, Marques–Bonet T, et al. 2009. Personalized copy number and segmental duplication maps using next–generation sequencing. Nat. Genetics 41: 1061–1067.

Ammiraju JSS, Lu F, Sanyal A, et al. 2008. Dynamic evolution of *Oryza* genomes is revealed by comparative genomic analysis of a genus–wide vertical data set. Plant Cell 20:3191–3209.

Ansorge WJ. 2009. Next–generation DNA sequencing techniques. Nat. Biotechnol. 25:195–203.

Arendt D. 2003. Evolution of eyes and photoreceptor cell types. Int. J. Dev. Biol. 47:563–71.

Arthan W, McKain MR, Traiperm P, Welker CAD, Teisher JK, Kellogg EA. 2017. Phylogenomics of Andropogoneae (Panicoideae: Poaceae) of mainland Southeast Asia. Syst. Bot. 42:418–431.

Atkinson RRL, Mockford EJ, Bennett C, Christin P–A, Spriggs EL, Freckleton RP, Thompson K, Rees M, Osborne CP. 2016. C_4 photosynthesis boosts growth by altering physiology, allocation and size. Nat. Plants 2:16038.

Aubry S, Brown NJ, Hibberd JM. 2011. The role of proteins in C_3 plants prior to their recruitment into the C_4 pathway. J. Exp. Bot. 62:3049–3059.

Aubry S, Kelly S, Kümpers BMC, Smith–Unna RD, Hibberd JM. 2014. Deep Evolutionary Comparison of Gene Expression Identifies Parallel Recruitment of Trans–Factors in Two Independent Origins of C_4 Photosynthesis. PLoS Genet. 10:e1004365.

Barve A, Wagner A. 2013. A latent capacity for evolutionary innovation through exaptation in metabolic systems. Nature 500:203–206.

Bauwe H, Hagemann M, Fernie AR. 2010. Photorespiration: players, partners and origin. Trends Plant Sci. 15:330–336.

Bellasio C, Griffiths H. 2014. The operation of two decarboxylases, transamination, and partitioning of C_4 metabolic processes between mesophyll and bundle sheath cells allows light capture to be balanced for the maize C_4 pathway. Plant Physiol. 164:466–480.

Bellos E, Johnson MR, Coin LJM. 2012. cnvHiTSeq: integrative models for high–resolution copy number variation detection and genotyping using population sequencing data. Genome Biology 13:R120.

Benjamini Y, Speed TP. 2012. Summarizing and correcting the GC content bias in high–throughput sequencing. Nucleic Acids Research 40:e72.

Bennett MD, Smith JB. 1991. Nuclear DNA amounts in angiosperms. Philos. Trans. R. Soc. B Biol. Sci. 334:309–345.

Bennetzen JL, Schmutz J, Wang H, et al. 2012. Reference genome sequence of the model plant *Setaria*. Nat. Biotech. 30:555–561.

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456:53–59.

Besnard G, Muasya AM, Russier F, Roalson EH, Salamin N, Christin P–A. 2009. Phylogenomics of C_4 Photosynthesis in Sedges (Cyperaceae): Multiple Appearances and Genetic Convergence. Mol. Biol. Evol. 26:1909–1919.

Besnard G, Christin P–A, Malé P–JG, Coissac E, Ralimanana H, Vorontsova MS. 2013. Phylogenomics and taxonomy of Lecomtelleae (Poaceae), an isolated panicoid lineage from Madagascar. Ann. Bot. 112:1057–1066.

Besnard G, Christin P–A, Malé P–JG, Lhuillier E, Lauzeral C, Coissac E, Vorontsova MS. 2014. From museums to genomics: old herbarium specimens shed light on a C_3 to C_4 transition. J. Exp. Bot. 65:6711–6721.

Besnard G, Bianconi ME, Hackel J, Manzi S, Vorontsova MS, Christin P–A. 2018. Herbarium genomics retraces the origins of C_4 –specific carbonic anhydrase in Andropogoneae (Poaceae). Bot. Lett.:1–15.

Beuning KRM, Zimmerman KA, Ivory SJ, Cohen AS. 2011. Vegetation response to glacial–interglacial climate variability near Lake Malawi in the southern African tropics. Palaeogeography Palaeoclimatology Palaeoecology, 303:81–92.

Bivand R, Keitt T, Rowlingson B. 2017. rgdal: bindings for the geospatial data abstraction library. R package version 1.2–16.

Bläsing OE, Ernst K, Streubel M, Westhoff P, Svensson P. 2002. The non–photosynthetic phosphoenolpyruvate carboxylases of the C_4 dicot *Flaveria trinervia* – implications for the evolution of C_4 photosynthesis. Planta 215:448–456.

Bläsing OE, Westhoff P, Svensson P. 2000. Evolution of C_4 phosphoenolpyruvate carboxylase in *Flaveria*, a conserved serine residue in the carboxyl–terminal part of the enzyme is a major determinant for C_4–specific characteristics. J. Biological Chemistry 275:27917–27923.

Bock DG, Kane NC, Ebert DP, Rieseberg LH. 2014. Genome skimming reveals the origin of the Jerusalem Artichoke tuber crop species: Neither from Jerusalem nor an artichoke. New Phytol. 201:1021–1030.

Bomblies K, Doebley JF. 2005. Molecular evolution of FLORICAULA/LEAFY orthologs in the Andropogoneae (Poaceae). Mol. Biol. Evol. 22:1082–1094.

Bond JE, Garrison NL, Hamilton CA, Godwin RL, Hedin M, Agnarsson I. 2014. Phylogenomics resolves a spider backbone phylogeny and rejects a prevailing paradigm for orb web evolution. Curr. Biol. 24:1765–1771.

Bond WJ, Midgley GF, Woodward FI. 2003. What controls South African vegetation – climate or fire? South Afr. J. Bot. 69:79–91.

Bond WJ, Silander JA, Ranaivonasy J, Ratsirarson J. 2008. The antiquity of Madagascar's grasslands and the rise of C_4 grassy biomes. J. Biogeogr. 35:1743–1758.

Bor NL. 1955. Notes on Asiatic grasses: XXIII. Jansenella Bor, a new genus of Indian grasses. Kew Bull. 10:93.

Borba AR, Serra TS, Górska A, Gouveia P, Cordeiro AM, Reyna–Llorens I, Kneřová J, Barros PM, Abreu IA, Oliveira MM, Hibberd JM, Saibo NJM. 2018. Synergistic binding of bHLH transcription factors to the promoter of the maize NADP–ME gene used in C_4 photosynthesis is based on an ancient code found in the ancestral C_3 state. Mol. Biol. Evol. 35:1–40.

Botha CEJ. 1992. Plasmodesmatal distribution, structure and frequency in relation to assimilation in C_3 and C_4 grasses in southern Africa. Planta. 187:348–358.

Bouchenak–Khelladi Y, Slingsby JA, Verboom GA, Bond WJ. 2014. Diversification of C_4 grasses (Poaceae) does not coincide with their ecological dominance. Am. J. Bot. 101:300–307.

Bräutigam A, Gowik U. 2016. Photorespiration connects C_3 and C_4 photosynthesis. J. Exp. Bot. 67:2953–2962.

Brautigam A, Kajala K, Wullenweber J, Sommer M, Gagneul D, Weber KL, Carr KM, Gowik U, Maß J, Lercher MJ, Westhoff P, Hibberd JM, Weber APM. 2011. An mRNA Blueprint for C_4 Photosynthesis Derived from Comparative Transcriptomics of Closely Related C_3 and C_4 species. Plant Physiol. 155:142–156.

Brown DD, Wensink PC, Jordan E. 1972. A comparison of the ribosomal DNA's of *Xenopus laevis* and *Xenopus mulleri*: the evolution of tandem genes. J. Mol. Biol. 63:57–73.

Brown NJ, Newell CA, Stanley S, Chen JE, Perrin AJ, Kajala K, Hibberd JM. 2011. Independent and Parallel Recruitment of Preexisting Mechanisms Underlying C_4 Photosynthesis. Science 331:1436–1439.

Buerkle CA, Gompert Z. 2013. Population genomics based on low coverage sequencing: How low should we go? Mol. Ecol. 22:3028–3035.

Burke AC, Nelson CE, Morgan BA, Tabin C. 1995. Hox genes and the evolution of vertebrate axial morphology. Development. 121:333–346.

Burke SV, Wysocki WP, Zuloaga FO, Craine JM, Pires JC, Edger PP, Mayfield–Jones D, Clark LG, Kelchner SA, Duvall MR. 2016. Evolutionary relationships in panicoid grasses based on plastome phylogenomics (Panicoideae; Poaceae). BMC Plant Biol. 16:140.

Capella–Gutiérrez S, Silla–Martínez JM, Gabaldón T. 2009. trimAl: A tool for automated alignment trimming in large–scale phylogenetic analyses. Bioinformatics. 25:1972–1973.

Castoe TA, de Koning APJ, Kim H–M, Gu W, Noonan BP, Naylor G, Jiang ZJ, Parkinson CL, Pollock DD. 2009. Evidence for an ancient adaptive episode of convergent molecular evolution. Proc. Natl. Acad. Sci. 106:8986–8991.

Cerling TE. 1999. Paleorecords of C_4 plants and ecosystems. In: Sage RF, Monson RK, editors. C_4 Plant Biology. Academic Press.

Cerling TE, Haris JM, MacFadden BJ, et al. 1997. Global vegetation change through the Miocene/Pliocene boundary. Nature 89:153–158.

Chen L, DeVries AL, Cheng CH. 1997. Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. Proc. Natl. Acad. Sci. 94:3817–22.

Chen Z, Cheng C–H.C, Zhang J, Cao L, Chen L, Zhou L, Jin Y, Ye H, Deng C, Dai Z, Xu Q, Hu P, Sun S, Shen Y, Chen L. 2008. Transcriptomic and genomic evolution under constant cold in Antarctic notothenioid fish. Proc. Natl. Acad. Sci. 105:12944–9.

Christin P–A, Salamin N, Savolainen V, Duvall MR, Besnard G. 2007. C_4 Photosynthesis Evolved in Grasses via Parallel Adaptive Genetic Changes. Curr. Biol. 17:1241–1247.

Christin P–A, Besnard G, Samaritani E, Duvall MR, Hodkinson TR, Savolainen V, Salamin N. 2008. Oligocene CO_2 Decline Promoted C_4 Photosynthesis in Grasses. Curr. Biol. 18:37–43.

Christin P–A, Besnard G. 2009a. Two independent C_4 origins in Aristidoideae (Poaceae) revealed by the recruitment of distinct phosphoenolpyruvate carboxylase genes. Am. J. Bot. 96:2234–2239.

Christin P–A, Petitpierre B, Salamin N, Büchi L, Besnard G. 2009b. Evolution of C_4 phosphoenolpyruvate carboxykinase in grasses, from genotype to phenotype. Mol. Biol. Evol. 26:357–365.

Christin P–A, Samaritani E, Petitpierre B, Salamin N, Besnard G. 2009c. Evolutionary insights on C_4 photosynthetic subtypes in grasses from genomics and phylogenetics. Genome Biol. Evol. 1:221–230.

Christin P–A, Freckleton RP, Osborne CP. 2010a. Can phylogenetics identify C_4 origins and reversals? Trends Ecol. Evol. 25:403–409.

Christin P–A, Weinreich DM, Besnard G. 2010b. Causes and evolutionary significance of genetic convergence. Trends Genet. 26:400–405.

Christin P–A, Sage TL, Edwards EJ, Ogburn RM, Khoshravesh R, Sage RF. 2011. Complex evolutionary transitions and the significance of C_3–C_4 intermediate forms of photosynthesis in Molluginaceae. Evolution 65:643–660.

Christin P–A, Edwards EJ, Besnard G, Boxall SF, Gregory R, Kellogg EA, Hartwell J, Osborne CP. 2012a. Adaptive evolution of C_4 photosynthesis through recurrent lateral gene transfer. Curr. Biol. 22:445–449.

Christin P–A, Besnard G, Edwards EJ, Salamin N. 2012b. Effect of genetic convergence on phylogenetic inference. Mol. Phylogenet. Evol. 62:921–927.

Christin P–A, Wallace MJ, Clayton H, Edwards EJ, Furbank RT, Hattersley PW, Sage RF, MacFarlane TD, Ludwig M. 2012c. Multiple photosynthetic transitions, polyploidy, and lateral gene transfer in the grass subtribe Neurachninae. J. Exp. Bot. 63:6297–6308.

Christin P–A, Boxall SF, Gregory R, Edwards EJ, Hartwell J, Osborne CP. 2013. Parallel recruitment of multiple genes into C_4 photosynthesis. Genome Biol. Evol. 5:2174–2187.

Christin P–A, Osborne C. 2013. The recurrent assembly of C_4 photosynthesis, an evolutionary tale. Photosynth. Res. 117:163–175.

Christin P–A, Osborne CP, Chatelet DS, Columbus JT, Besnard G, Hodkinson TR, Garrison LM, Vorontsova MS, Edwards EJ. 2013. Anatomical enablers and the evolution of C_4 photosynthesis in grasses. Proc. Natl. Acad. Sci. 110:1381–1386.

Christin P–A, Osborne CP. 2014. The evolutionary ecology of C_4 plants. New Phytol. 204:765–781.

Christin P–A, Spriggs E, Osborne CP, Strömberg CAE, Salamin N, Edwards EJ. 2014. Molecular dating, evolutionary rates, and the age of the grasses. Syst. Biol. 63:153–165.

Christin P–A, Arakaki M, Osborne CP, Edwards EJ. 2015. Genetic enablers underlying the clustered evolutionary origins of C_4 photosynthesis in angiosperms. Mol. Biol. Evol. 32:846–858.

Clark CJ, McGuire JA, Bonaccorso E, Berv JS, Prum RO. 2018. Complex coevolution of wing, tail, and vocal sounds of courting male bee hummingbirds. Evolution 72:630–646.

Clark J, Hidalgo O, Pellicer J, et al. 2016. Genome evolution of ferns: evidence for relative stasis of genome size across the fern phylogeny. New Phytol., 210:1072–1082.

Clayton WD. 1979. Notes on *Setaria* (Gramineae). Kew Bulletin 33:501–509.

Clayton WD, Renvoize SA. 1986. Genera graminum. Grasses of the World. Genera graminum. Grasses of the World 13.

Clayton WD, Govaerts R, Harman KT, Williamson H, Vorontsova MS. 2016. World checklist of Poaceae, facilitated by the Royal Botanic Gardens, Kew. Available from http://apps.kew.org/wcsp/.

Clift PD. 2006. Controls on the erosion of Cenozoic Asia and the flux of clastic sediment to the ocean. Earth Planet. Sci. Lett. 241:571–580.

Clift PD, Hodges KV, Heslop D, Hannigan R, Long HV, Calves G. 2008. Correlation of Himalayan exhumation rates and Asian monsoon intensity. Nat. Geosci. 1:875–880.

Cohen AS, Stone JR, Beuning KR, et al. 2007. Ecological consequences of early late pleistocene megadroughts in tropical Africa. Proc. Natl. Acad. Sci. 104:16422–16427.

Cohen S, Houben A, Segal D. 2007. Extrachromosomal circular DNA derived from tandemly repeated genomic sequences in plants. Plant J. 53:1027–1034.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: how duplicated genes find new functions. Nat. Rev. Genet. 9:938–50.

Conant GC, Birchler JA, Pires JC. 2014. Dosage, duplication, and diploidization: Clarifying the interplay of multiple models for duplicate gene evolution over time. Curr. Opin. Plant Biol. 19:91–98.

Cook DE, Lee TG, Guo XL, Melito S, Wang K, Bayless AM, Wang JP, Hughes TJ, Willis DK, Clemente TE, Diers BW, Jiang JM, Hudson ME, Bent AF. 2012. Copy Number Variation of Multiple Genes at Rhg1 Mediates Nematode Resistance in Soybean. Science 338:1206–1209.

Cotton JL, Wysocki WP, Clark LG, Kelchner SA, Pires JC, Edger PP, Mayfield–Jones D, Duvall MR. 2015. Resolving deep relationships of PACMAD grasses: a phylogenomic approach. BMC Plant Biol. 15:178.

Covshoff S, Hibberd JM. 2012. Integrating C_4 photosynthesis into C_3 crops to increase yield potential. Curr. Opin. Biotechnol. 23:209–214.

Cunningham CW, Omland KE, Oakley TH. 1998. Reconstructing ancestral character states: a critical reappraisal. Trends Ecol. Evol. 13:361–366.

Damen WGM, Saridaki T, Averof M. 2002. Diverse Adaptations of an Ancestral Gill: A Common Evolutionary Origin for Wings, Breathing Organs, and Spinnerets. Curr. Biol. 12:1711–1716.

Danila FR, Quick WP, White RG, Furbank RT, von Caemmerer S. 2016. The Metabolite Pathway between Bundle Sheath and Mesophyll: Quantification of Plasmodesmata in Leaves of C_3 and C_4 Monocots. Plant Cell 28:1461–1471.

Danila FR, Quick WP, White RG, Kelly S, von Caemmerer S, Furbank RT. 2018. Multiple mechanisms for enhanced plasmodesmata density in disparate subtypes of C_4 grasses. J. Exp. Bot. 69:1135–1145.

Darwin C. 1859. On the origins of species by means of natural selection. London: Murray.

Dengler NG, Dengler RE, Hattersley PW. 1985. Differing ontogenetic origins of PCR ("Kranz") sheaths in leaf blades of C_4 grasses (Poaceae). Am. J. Bot. 72:284–302.

Dengler NG, Donnelly PM, Dengler RE. 1996. Differentiation of bundle sheath, mesophyll, and distinctive cells in the C_4 grass *Arundinella hirta* (Poaceae). Am. J. Bot. 83:1391–1405.

Dengler NG, Nelson T. 1999. Leaf structure and development in C_4 plants. In: Sage RF, Monson RK, editors. C_4 Plant Biology. Academic Press. p:133–172.

Dengler RE, Dengler NG. 1990. Leaf vascular architecture in the atypical C_4 NADP – malic enzyme grass *Arundinella hirta*. Can. J. Bot. 68:1208–1221.

Dionne M, Caron F, Dodson JJ, Bernatchez L. 2008. Landscape genetics and hierarchical genetic structure in Atlantic salmon: the interaction of gene flow and local adaptation. Mol. Ecol. 17:2382–2396.

Dodsworth S. 2015. Genome skimming for next–generation biodiversity analysis. Trends Plant Sci. 20:525–527.

Dohm JC, Lottaz C, Borodina T, Himmelbauer H. 2008. Substantial biases in ultra–short read data sets from high–throughput DNA sequencing. Nucleic Acids Research 36:e105.

Doležel J, Sgorbati S, Lucretti S. 1992. Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. Physiologia Plantarum, 85:625–631.

Doležel J, Greilhuber J, Suda J. 2007. Estimation of nuclear DNA content in plants using flow cytometry. Nat. Protocols, 2:2233–2244.

Doust AN, Penly AM, Jacobs SWL, Kellogg EA. 2007. Congruence, conflict, and polyploidization shown by nuclear and chloroplast markers in the monophyletic “bristle clade” (Paniceae, Panicoideae, Poaceae). Syst. Bot. 32:531–544.

Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:699–710.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Dunning LT, Lundgren MR, Moreno–Villena JJ, Namaganda M, Edwards EJ, Nosil P, Osborne CP, Christin P–A. 2017a. Introgression and repeated co–option facilitated the recurrent emergence of C_4 photosynthesis among close relatives. Evolution 71:1541–1555.

Dunning LT, Liabot A–L, Olofsson JK, Smith EK, Vorontsova MS, Besnard G, Simpson KJ, Lundgren MR, Addicott E, Gallagher RV, Chu Y, Pennington RT, Christin P–A, Lehmann CER. 2017b. The recent and rapid spread of Themeda triandra. Bot. Lett. 164:327–337.

Durand EY, Patterson N, Reich D, Slatkin M. 2011. Testing for ancient admixture between closely related populations. Mol. Biol. Evol. 28:2239–2252.

Duvall MR, Saar DE, Grayburn WS, Holbrook GP. 2003. Complex Transitions between C_3 and C_4 Photosynthesis during the Evolution of Paniceae: A Phylogenetic Case Study Emphasizing the Position of Steinchisma hians (Poaceae), a C_3-C_4 Intermediate. Int. J. Plant Sci. 164:949–958.

Earl DA, vonHoldt BM. 2012. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. Conservation Genetics Resources, 4:359–361.

Ebihara A, Ishikawa H, Matsumoto S, et al. 2005. Nuclear DNA, chloroplast DNA, and ploidy analysis clarified biological complexity of the Vandenboschia radicans complex (Hymenophyllaceae) in Japan and adjacent areas. American J. Botany, 92:1535–1547.

Edgar RC. 2004. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

Edwards EJ, Still CJ. 2008. Climate, phylogeny and the ecological distribution of C_4 grasses. Ecol. Lett. 11:266–276.

Edwards EJ, Osborne CP, Stromberg CAE, Smith SA, Bond WJ, Christin P–A, Cousins AB, Duvall MR, Fox DL, Freckleton RP, Ghannoum O, Hartwell J, Huang Y, Janis CM, Keeley JE, Kellogg EA, Knapp AK, Leakey ADB, Nelson DM, Saarela JM, Sage RF, Sala O.E, Salamin N, Still CJ, Tipple B. 2010. The origins of C_4 grasslands: integrating evolutionary and ecosystem science. Science 328:587–591.

Edwards EJ, Smith SA. 2010. Phylogenetic analyses reveal the shady history of C_4 grasses. Proc. Natl. Acad. Sci. 107:2532–2537.

Edwards EJ, Ogburn RM. 2012. Angiosperm Responses to a Low–CO_2 World: CAM and C_4 Photosynthesis as Parallel Evolutionary Trajectories. Int. J. Plant Sci. 173:724–733.

Ehleringer JR, Bjorkman O. 1977. Quantum Yields for CO_2 Uptake in C_3 and C_4 Plants: Dependence on Temperature, CO_2 and O_2 Concentration. Plant Physiol. 59:86–90.

Ehleringer JR, Sage RF, Flanagan LB, Pearcy RW. 1991. Climate change and the evolution of C_4 photosynthesis. Trends Ecol. Evol. 6:95–99.

Ehleringer JR, Cerling TE, Helliker BR. 1997. C_4 photosynthesis, atmospheric CO_2 and climate. Oecologia 112:285–299.

Ekblom R, Galindo J. 2011. Applications of next generation sequencing in molecular ecology of non–model organisms. Heredity 107:1–15.

Eldredge N. and Gould SJ. 1972. Punctuated equilibria: an alternative to phyletic gradualism. In Models in Paleobiology (Schopf, T, ed.), pp. 82–115:Freeman Cooper.

Eldredge N, Gould SJ, Coyne JA, Charlesworth B. 1997. On Punctuated Equilibria. Science 276:337c–341.

Elena SF, Cooper VS, Lenski RE. 1996. Punctuated evolution caused by selection of rare beneficial mutations. Science 272:1802–4.

Ellis R. 1974. Significance of the occurrence of both Kranz and non–Kranz leaf anatomy in the grass species *Alloteropsis semialata*. South African J. Science 70:169–173.

Ellstrand NC. 2014. Is gene flow the most important evolutionary force in plants? American J. Botany 101:737–753.

Emms DM, Covshoff S, Hibberd JM, 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. Mol. Biol. Evol. 33:1796–1806.

Endress PK. 2011. Evolutionary diversification of the flowers in angiosperms. Am. J. Bot. 98:370–396.

Estep MC, Diaz DMV, Zhong J, Kellogg EA. 2012. Eleven diverse nuclear–encoded phylogenetic markers for the subfamily Panicoideae (Poaceae). Am. J. Bot. 99:e443–e446.

Estep MC, McKain MR, Vela Diaz D, Zhong J, Hodge JG, Hodkinson TR, Layton DJ, Malcomber ST, Pasquet R, Kellogg EA. 2014. Allopolyploidy, diversification, and the Miocene grassland expansion. Proc. Natl. Acad. Sci. 111:15149–15154.

Evanno G, Regnaut S, Goudet J. 2005. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. Mol. Ecol. 14:2611–2620.

Felsenstein J. 2003. Inferring Phylogenies. Sunderland (MA): Sinauer Associates.

Fisher AE, McDade LA, Kiel CA, et al.. 2015. Evolutionary history of *Blepharis* (Acanthaceae) and the origin of C_4 photosynthesis in section Acanthodium. International J. Plant Sciences 176:770–790.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, al. et. 1995. Whole–genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science 269:496–512.

Forrestel EJ, Donoghue MJ, Smith MD. 2014. Convergent phylogenetic and functional responses to altered fire regimes in mesic savanna grasslands of North America and South Africa. New Phytol. 203:1000–1011.

Fournier–Level A, Korte A, Cooper MD, Nordborg M, Schmitt J, Wilczek AM. 2011. A map of local adaptation in *Arabidopsis thaliana*. Science 334:86–9.

Fumagalli M, Vieira FG, Korneliussen TS, et al.. 2013. Quantifying Population Genetic Differentiation from Next–Generation Sequencing Data. Genetics 195:979–992.

Furbank RT. 2011. Evolution of the C_4 photosynthetic mechanism: Are there really three C_4 acid decarboxylation types? J. Exp. Bot. 62:3103–3108.

Geber MA, Griffen LR. 2003. Inheritance and natural selection on functional traits. Int. J. Plant Sci. 164:S21–S42.

Gehring WJ. 2004. Historical perspective on the development and evolution of eyes and photoreceptors. Int. J. Dev. Biol. 48:707–17.

Gould SJ, Eldredge N. 1977. Punctuated equilibria: the tempo and mode of evolution reconsidered. Paleobiology. 3:115–151.

Gould SJ, Vrba ES. 1982. Exaptation—a missing term in the science of form. Paleobiology. 8:4–15.

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 16:1077–1090.

Gowik U, Bräutigam A, Weber KL, Weber APM, Westhoff P. 2011. Evolution of C_4 photosynthesis in the genus *Flaveria*: how many and which genes does it take to make C_4? Plant Cell 23:2087–2105.

Gowik U, Westhoff P. 2011. The Path from C_3 to C_4 Photosynthesis. Plant Physiol. 155:56–63.

Grass Phylogeny Working Group II (GPWG II). 2012. New grass phylogeny resolves deep evolutionary relationships and discovers C_4 origins. New Phytol. 193:304–312.

Green RE, Krause J, Briggs AW, et al. 2010. A draft sequence of the Neandertal genome. Science 328:710–722.

Griffith DM, Anderson TM, Osborne CP, Strömberg CAE, Forrestel EJ, Still CJ. 2015. Biogeographically distinct controls on C_3 and C_4 grass distributions: Merging community and physiological ecology. Global Ecology and Biogeography, 24:304–313.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum–likelihood phylogenies: Assessing the performance of PhyML 3.0. Syst. Biol. 59:307–321.

Guo ZT, Ruddiman WF, Hao QZ, Wu HB, Qiao YS, Zhu RX, Peng SZ, Wel JJ, Yuan BY, Llu TS. 2002. Onset of Asian desertification by 22 Myr ago inferred from loess deposits in China. Nature 416:159–163.

Gutierrez M, Gracen VE, Edwards GE. 1974. Biochemical and cytological relationships in C_4 plants. Planta 119:279–300.

Haberlandt G. 1884. Physiologische Pflanzenanatomie. Leipzig: Engelmann.

Hackel J, Vorontsova MS, Nanjarisoa O.P, Hall RC, Razanatsoa J, Malakasi P, Besnard G. 2018. Grass diversification in Madagascar: In situ radiation of two large C_3 shade clades and support for a Miocene to Pliocene origin of C_4 grassy biomes. J. Biogeogr. 45:750–761.

Hackl T, Hedrich R, Schultz J, Förster F. 2014. proovread : large–scale high–accuracy PacBio correction through iterative short read consensus. Bioinformatics. 30:3004–3011.

Hall LN, Rossini L, Cribb L, Langdale JA. 1998. GOLDEN 2: A Novel Transcriptional Regulator of Cellular Differentiation in the Maize Leaf. Plant Cell Online. 10:925–936.

Hartl DL, Clark AG. 2007. Principles of population genetics. Sunderland (MA) Sinauer Associates.

Hartley W. 1958A. Studies on the origin, evolution and distribution of the Gramineae. I. The tribe Andropogoneae. Aust. J. Bot. 6:116–128.

Hartley W. 1958b. Studies on the origin, evolution, and distribution of the Gramineae. II. The tribe Paniceae. Aust. J. Bot. 6:343.

Hartley W, Slater C. 1960. Studies on the origin, evolution, and distribution of the Gramineae. III. The tribes of the subfamily Eragrostoideae. Aust. J. Bot. 8:256.

Hartley W. 1961. Studies on the origin, evolution and distribution of the Gramineae. IV. The genus Poa L. Aust. J. Bot. 9:152.

Hartley W. 1973. Studies on the Origin, Evolution, and Distribution of the Gramineae. V. The Subfamily Festucoideae. Aust. J. Bot. 21:201.

Hasegawa M, Kishino H. 1989. Heterogeneity of tempo and mode of mitochondrial DNA evolution among mammalian orders. Japanese J. Genet. 64:243–258.

Hatch MD. 1987. C_4 photosynthesis: a unique blend of modified biochemistry, anatomy and ultrastructure. Biochim. Biophys. Acta 895:81–106.

Hatch MD, Slack CR. 1966. Photosynthesis by sugar–cane leaves. A new carboxylation reaction and the pathway of sugar formation. Biochem. J. 101:103–11.

Hattersley PW. 1984. Characterization of C_4 type leaf anatomy in grasses (Poaceae). Mesophyll: bundle sheath area ratios. Ann. Bot. 53:163–180.

Heckmann D, Schulze S, Denton A, Gowik U, Westhoff P, Weber APM, Lercher MJ. 2013. Predicting C_4 photosynthesis evolution: Modular, individually adaptive steps on a mount fuji fitness landscape. Cell 153:1579–1588.

Heckmann D. 2016. C_4 photosynthesis evolution: The conditional Mt. Fuji. Curr. Opin. Plant Biol. 31:149–154.

Hedrick PW. 2013. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. Mol. Ecol. 22:4606–4618.

Hewitt G. 2000. The genetic legacy of the Quaternary ice ages. Nature 405:907–913.

Hibberd JM, Covshoff S. 2010. The Regulation of Gene Expression Required for C_4 Photosynthesis. Annu. Rev. Plant Biol. 61:181–207.

Hilu KW, Alice LA, Liang H. 1999. Phylogeny of Poaceae inferred from matK sequences. Ann. Miss. Bot. Gar. 86:835–851.

Hohmann–Marriott MF, Blankenship RE. 2011. Evolution of Photosynthesis. Annu. Rev. Plant Biol. 62:515–548.

Hotopp JCD, Clark ME, Oliveira DCSG, Foster JM, Fischer P, Muñoz Torres MC, Giebel JD, Kumar N, Ishmael N, Wang S, Ingram J, Nene R. V, Shepard J, Tomkins J, Richards S, Spiro DJ, Ghedin E, Slatko BE, Tettelin H, Werren JH. 2007. Widespread lateral gene transfer from intracellular bacteria to multicellular eukaryotes. Science 317:1753–6.

Hsu CC. 1975. Taiwan Grasses: by Chien–chang Hsu. Taiwan Provincial Education Association, Taipei.

Huang C–F, Yu C–P, Wu Y–H, Lu M–YJ, Tu S–L, Wu S–H, Shiu S–H, Ku MSB, Li W–H. 2017. Elevated auxin biosynthesis and transport underlie high vein density in C_4 leaves. Proc. Natl. Acad. Sci. 114:E6884–E6891.

Huang P, Studer AJ, Schnable JC, Kellogg EA, Brutnell TP. 2017. Cross species selection scans identify components of C_4 photosynthesis in the grasses. J. Exp. Bot. 68:127–135.

Huerta–Sánchez E, Jin X, Asan, Bianba Z, Peter B.M, Vinckenbosch N, Liang Y, Yi X, He M, Somel M, Ni P, Wang B, Ou X, Huasang, Luosang J, Cuo ZXP, Li K, Gao G, Yin Y, Wang W, Zhang X, Xu X, Yang H, Li Y, Wang J, Wang J, Nielsen R. 2014. Altitude adaptation in Tibetans caused by introgression of Denisovan–like DNA. Nature 512:194–197.

Hufford MB, Lubinsky P, Pyhäjärvi T, et al. 2013. The genomic signature of crop–wild introgression in maize. Plos Genetics 9:e1003477.

Hughes CL, Kaufman TC. 2002. Hox genes and the evolution of the arthropod body plan1. Evol. Dev. 4:459–499.

Hulbert LC. 1988. Causes of fire effects in tallgrass prairie. Ecology 69:46–58.

Ibrahim DG, Gilbert ME, Ripley BS, Osborne CP. 2008. Seasonal differences in photosynthesis between the C_3 and C_4 subspecies of *Alloteropsis semialata* are offset by frost and drought. Plant. Cell Environ. 31:1038–1050.

Ibrahim DG, Burke T, Ripley BS, Osborne CP. 2009. A molecular phylogeny of the genus *Alloteropsis* (Panicoideae, Poaceae) suggests an evolutionary reversion from C_4 to C_3 photosynthesis. Annals of Botany, 103:127–136.

International Human Genome Sequencing Consortium 2001. Initial sequencing and analysis of the human genome. Nature 409:860–921.

International Human Genome Sequencing Consortium 2004. Finishing the euchromatic sequence of the human genome. Nature 431:931–945.

Jakobsson M, Rosenberg NA. 2007. CLUMPP: A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics 23:1801–1806.

John CR, Smith–Unna RD, Woodfield H, Covshoff S, Hibberd JM. 2014. Evolutionary Convergence of Cell–Specific Gene Expression in Independent Lineages of C_4 Grasses. Plant Physiol. 165:62–75.

Jones S, Burke S, Duvall M. 2014. Phylogenomics, molecular evolution, and estimated ages of lineages from the deep phylogeny of Poaceae. Plant Syst. Evol. 300:1421–1436.

Kadereit G, Borsch T, Weising K, Freitag H. 2003. Phylogeny of Amaranthaceae and Chenopodiaceae and the Evolution of C_4 Photosynthesis. Int. J. Plant Sci. 164:959–986.

Kadereit G, Bohley K, Lauterbach M, Tefarikis DT, Kadereit JW. 2017. C_3–C_4 intermediates may be of hybrid origin – a reminder. New Phytol. 215:70–76.

Kaessmann H, Vinckenbosch N, Long M. 2009. RNA–based gene duplication: mechanistic and evolutionary insights. Nat. Reviews: Genetics 10:19–31.

Kajala K, Covshoff S, Karki S, Woodfield H, Tolley BJ, Dionora MJA, Mogul RT, Mabilangan AE, Danila FR, Hibberd JM, Quick WP. 2011. Strategies for engineering a two–celled C_4 photosynthetic pathway into rice. J. Exp. Bot. 62:3001–3010.

Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. 2017. ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat. Methods 14:587–589.

Kanai R, Edwards GE. 1999. The biochemistry of C_4 photosynthesis. In: Sage RF, Monson RK, editors. C_4 Plant Biology. Academic Press. p. 49–87.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol. Biol. Evol. 30:772–780.

Kearse M, Moir R, Wilson A, Stones–Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, Thierer T, Ashton B, Meintjes P, Drummond A. 2012. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649.

Keeley JE, Rundel PW. 2003. Evolution of CAM and C_4 Carbon-Concentrating Mechanisms. Int. J. Plant Sci. 164:S55–S77.

Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet. 9:605–618.

Kellogg EA. 1999. Phylogenetic Aspects of the Evolution of C_4 Photosynthesis. In: Sage RF, Monson RK, editors. C_4 Plant Biology. Academic Press. p. 411–444.

Kellogg EA. 2015. Flowering Plants. Monocots: Poaceae. Heidelberg: Springer. 416p.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV. 2002. Selection in the evolution of gene duplications. Genome Biology 3:1–9.

Kondrashov FA. 2012. Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc. R. Soc. B Biol. Sci. 279:5048–5057.

Korneliussen TS, Albrechtsen A, Nielsen R et al. 2014. ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics 15:356.

Kortschak HP, Hartt CE, Burr GO. 1965. Carbon Dioxide Fixation in Sugarcane Leaves. Plant Physiol. 40:209–13.

Kreiner JM, Stinchcombe JR, Wright SI. 2018. Population Genomics of Herbicide Resistance: Adaptation via Evolutionary Rescue. Annu. Rev. Plant Biol. 69:611–635.

Kroymann J, Mitchell–Olds T. 2005. Epistasis and balanced polymorphism influencing complex trait variation. Nature 435:95–98.

Ku MSB, Monson RK, Littlejohn RO, Nakamoto H, Fisher DB, Edwards GE. 1983. Photosynthetic Characteristics of C_3–C_4 Intermediate *Flaveria* Species : I. Leaf Anatomy,

Photosynthetic Responses to O2 and CO_2, and Activities of Key Enzymes in the C_3 and C_4 Pathways. Plant Physiol. 71:944–948.

Kukalová–Peck J, Lawrence JF. 1993. Evolution of the hind wing in Coleoptera. Can. Entomol. 125:181–258.

Külahoglu C, Denton AK, Sommer M, Maß J, Schliesky S, Wrobel TJ, Berckmans B, Gongora–Castillo E, Buell CR, Simon R, De Veylder L, Bräutigam A, Weber APM. 2014. Comparative transcriptome atlases reveal altered gene expression modules between two Cleomaceae C_3 and C_4 plant species. Plant Cell 26:3243–60.

Kumar S. 2005. Molecular clocks: four decades of evolution. Nat. Rev. Genet. 6:654–662.

Lamb TD, Collin SP, Pugh EN. 2007. Evolution of the vertebrate eye: opsins, photoreceptors, retina and eye cup. Nat. Rev. Neurosci. 8:960–976.

Lanciano S, Carpentier M–C, Llauro C, Jobet E, Robakowska–Hyzorek D, Lasserre E, Ghesquière A, Panaud O, Mirouze M. 2017. Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants. PLOS Genet. 13:e1006630.

Langdale JA. 2011. C_4 Cycles: Past, Present, and Future Research on C_4 Photosynthesis. Plant Cell 23:3879–3892.

Langley CH, Fitch WM. 1974. An examination of the constancy of the rate of molecular evolution. J. Mol. Evol. 3:161–177.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory–efficient alignment of short DNA sequences to the human genome. Genome Biology, 10:R25.

Langmead B, Salzberg SL. 2012. Fast gapped–read alignment with Bowtie 2. Nat. Methods 9:357–359.

Leegood RC. 2008. Roles of the bundle sheath cells in leaves of C_3 plants. J. Exp. Bot. 59:1663–1673.

Leinonen T, Cano JM, Makinen H, Merila J. 2006. Contrasting patterns of body shape and neutral genetic divergence in marine and lake populations of threesping sticklebacks. J. Evolutionary Biology, 19:1803–1812.

Lenski RE, Ofria C, Pennock RT, Adami C. 2003. The evolutionary origin of complex features. Nature 423:139–144.

Li F–W, Villarreal JC, Kelly S, Rothfels CJ, Melkonian M, Frangedakis E, Ruhsam M, Sigel E.M, Der JP, Pittermann J, Burge DO, Pokorny L, Larsson A, Chen T, Weststrand S, Thomas P, Carpenter E, Zhang Y, Tian Z, Chen L, Yan Z, Zhu Y, Sun X, Wang J, Stevenson DW, Crandall–Stotler BJ, Shaw AJ, Deyholos MK, Soltis DE, Graham SW, Windham MD, Langdale JA, Wong GK–S, Mathews S, Pryer KM. 2014. Horizontal transfer of an adaptive chimeric photoreceptor from bryophytes to ferns. Proc. Natl. Acad. Sci. 111:6672–7.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079.

Li Y, Sidore C, Kang HM, Boehnke M, Abecasis GR. 2011. Low–coverage sequencing: implications for design of complex trait association studies. Genome Research, 21:940–951.

Liebenberg EJL, Fossey A. 2001. Comparative cytogenetic investigation of the two subspecies of the grass *Alloteropsis semialata* (Poaceae). Botanical J. the Linnean Society 137:243–248.

Linder HP, de Klerk HM, Born J, Burgess ND, Fjeldså J, Rahbek C. 2012. The partitioning of Africa: Statistically defined biogeographical regions in sub–Saharan Africa. J. Biogeography 39:1189–1205.

Linder HP, Lehmann CER, Archibald S, Osborne CP, Richardson DM. 2018. Global grass (Poaceae) success underpinned by traits facilitating colonization, persistence and habitat transformation. Biol. Rev. 93: 1125–1144.

Long SP. 1999. Environmental Responses. In: Sage RF, Monson RK, editors. C_4 Plant Biology. Academic Press. p. 215–249.

Loureiro J, Rodriguez E, Doležel J, Santos C. 2007. Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species. Annals of Botany, 100:875–888.

Lundgren MR, Osborne CP, Christin P–A. 2014. Deconstructing Kranz anatomy to understand C_4 evolution. J. Exp. Bot. 65:3357–3369.

Lundgren MR, Besnard G, Ripley BS, Lehmann CER, Chatelet DS, Kynast RG, Namaganda M, Vorontsova MS, Hall RC, Elia J, Osborne CP, Christin P–A. 2015. Photosynthetic innovation broadens the niche within a single species. Ecol. Lett. 18:1021–1029.

Lundgren MR, Christin P–A, Escobar E.G, Ripley BS, Besnard G, Long CM, Hattersley PW, Ellis RP, Leegood RC, Osborne CP. 2016. Evolutionary implications of C_3–C_4 intermediates in the grass *Alloteropsis semialata*. Plant, Cell Environ. 39:1874–1885.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu S–M, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam T–W, Wang J. 2012. SOAPdenovo2: an empirically improved memory–efficient short–read de novo assembler. Gigascience. 1:18.

Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155.

MacFadden BJ. 1986. Fossil horses from "Eohippus" (Hyracotherium) to Equus: scaling, Cope's Law, and the evolution of body size. Paleobiology. 12:355–369.

Mahelka V, Krak K, Kopecký D, Fehrer J, Šafář J, Bartoš J, Hobza R, Blavet N, Blattner FR. 2017. Multiple horizontal transfers of nuclear ribosomal genes between phylogenetically distinct grass lineages. Proc. Natl. Acad. Sci. 114.

Majeran W, Zybailov B, Ytterberg AJ, Dunsmore J, Sun Q, van Wijk KJ. 2008. Consequences of C_4 differentiation for chloroplast membrane proteomes in maize mesophyll and bundle sheath cells. Mol. Cell. Proteomics 7:1609–38.

Mallmann J, Heckmann D, Bräutigam A, Lercher MJ, Weber APM, Westhoff P, Gowik U. 2014. The role of photorespiration during the evolution of C_4 photosynthesis in the genus *Flaveria*. ELife 3:e02478.

Marek P.E, Moore W. 2015. Discovery of a glowing millipede in California and the gradual evolution of bioluminescence in Diplopoda. Proc. Natl. Acad. Sci. USA 112:6419–6424.

Martinson E.O, Mrinalini, Kelkar YD, Chang C–H, Werren JH. 2017. The Evolution of Venom by Co–option of Single–Copy Genes. Curr. Biol. 27:2007–2013.e8.

Mayr E. 1982. The growth of biological thought: diversity, evolution, and inheritance. Belknap Press of Harvard University Press, Cambridge, MA. 974 pp. Harvard University Press.

McGee MD, Borstein SR, Neches RY, Buescher HH, Seehausen O, Wainwright PC. 2015. A pharyngeal jaw evolutionary innovation facilitated extinction in Lake Victoria cichlids. Science 350:1077–1079.

McKown AD, Dengler NG. 2007. Key innovations in the evolution of Kranz anatomy and C_4 vein pattern in *Flaveria* (Asteraceae). American J. Botany, 94:382–399.

McKown AD, Moncalvo JM, Dengler NG. 2005. Phylogeny of *Flaveria* (Asteraceae) and inference of C_4 photosynthesis evolution. American J. Botany, 92:1911–1928.

McLennan D. 2008. The Concept of Co–option: Why Evolution Often Looks Miraculous. Evol. Educ. Outreach. 1:247–258.

Meléndez–Hevia E, Waddell TG, 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. J. Molecular Evolution, 43:293–303.

Messer PW, Ellner SP, Hairston NG. 2016. Can Population Genetics Adapt to Rapid Evolution? Trends Genet. 32:408–418.

Metcalfe CR. 1960. Anatomy of the monocotyledons – I. Gramineae. Oxford, 267p.

Metzger MJ, Paynter AN, Siddall ME, Goff SP. 2018. Horizontal transfer of retrotransposons between bivalves and other aquatic species of multiple phyla. Proc. Natl. Acad. Sci. 115:E4227–E4235.

Miller CT, Glazer AM, Summers BR, et al. 2014. Modular skeletal evolution in sticklebacks is controlled by additive and clustered quantitative trait loci. Genetics 197:405–420.

Milligan BG. 1986. Punctuated Evolution Induced by Ecological Change. Am. Nat. 127:522–532.

Møller AP, Pomiankowski A. 1993. Punctuated Equilibria or Gradual Evolution: Fluctuating Asymmetry and Variation in the Rate of Evolution. J. Theor. Biol. 161:359–367.

Monson RK, Edwards GE, Ku MSB. 1984. C_3–C_4 Intermediate Photosynthesis in Plants. Bioscience. 34:563–574.

Monson RK, Teeri JA, Ku MS, Gurevitch J, Mets LJ, Dudley S. 1988. Carbon–isotope discrimination by leaves of *Flaveria* species exhibiting different amounts of C_3– and C_4–cycle co–function. Planta, 174:145–151.

Monson RK, Moore BD. 1989. On the significance of C_3–C_4 intermediate photosynthesis to the evolution of C_4 photosynthesis. Plant. Cell Environ. 12:689–699.

Monson RK. 1999. The origins of C_4 genes and evolutionary pattern in the C_4 metabolic phenotype. In: Sage RF, Monson RK, eds. C_4 plant biology. San Diego: Academic Press, 377–410.

Monson RK. 2003. Gene Duplication, neofunctionalization, and the evolution of C_4 photosynthesis. International J. Plant Sciences 164:S43–S54.

Moreno–Villena JJ, Dunning LT, Osborne CP, Christin P–A. 2018. Highly Expressed Genes Are Preferentially Co–Opted for C_4 Photosynthesis. Mol. Biol. Evol. 35:94–106.

Morjan CL, Rieseberg LH. 2004. How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. Mol. Ecol., 13:1341–1356.

Morrone O, Aliscioni S.S, Veldkamp JF, Pensiero JF, Zuloaga F.O, Kellogg EA. 2013. Revision of the Old World species of *Setaria* (Poaceae: Panicoideae: Paniceae). In: Systematic Botany Monographs, volume 96.

Mouchés C, Pasteur N, Berge J, Hyrien O, Raymond M, de Saint Vincent B, de Silvestri M, Georghiou G. 1986. Amplification of an esterase gene is responsible for insecticide resistance in a California Culex mosquito. Science 233:778–780.

Nadeau NJ, Jiggins CD. 2010. A golden age for evolutionary genetics? Genomic studies of adaptation in natural populations. Trends Genet. 26:484–492.

Nair VJ, Ramachandran VS, Sreekumar PV. 1982. Chandrasekharania: a new genus of Poaceae from Kerala, India. Proc. Indian Acad. Sci. – Sect. B. Part 3:Plant Sci. 91:79–82.

Nei M, Rooney AP. 2005. Concerted and birth–and–death evolution of multigene families. Ann. Rev. Genetics 39:121–52.

Nelson–Sathi S, Sousa FL, Roettger M, Lozada–Chávez N, Thiergart T, Janssen A, Bryant D, Landan G, Schönheit P, Siebers B, McInerney JO, Martin WF. 2015. Origins of major archaeal clades correspond to gene acquisitions from bacteria. Nature 517:77–80.

Nguyen LT, Schmidt H.A, Von Haeseler A, Minh BQ. 2015. IQ–TREE: A fast and effective stochastic algorithm for estimating maximum–likelihood phylogenies. Mol. Biol. Evol. 32:268–274.

Nicod J, Davies RW, Cai N, et al. 2016. Genome–wide association of multiple complex traits in outbred mice by ultra–low–coverage sequencing. Nat. Genetics 48:912–918.

O'Leary MH. 1988. Carbon isotopes in photosynthesis. BioScience 38:328–336.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304.

Ohno S. 1970. Evolution by Gene Duplication, Springer.

Osborne CP. 2008. Atmosphere, ecology and evolution: what drove the Miocene expansion of C_4 grasslands? J. Ecol. 96:35–45.

Osborne CP, Wythe EJ, Ibrahim DG, Gilbert ME, Ripley BS. 2008. Low temperature effects on leaf physiology and survivorship in the C_3 and C_4 subspecies of *Alloteropsis semialata*. J. Exp. Bot. 59:1743–1754.

Osborne CP, Freckleton RP. 2009. Ecological selection pressures for C_4 photosynthesis in the grasses. Proc. R. Soc. B Biol. Sci. 276:1753–1760.

Osborne CP, Sack L. 2012. Evolution of C_4 plants: a new hypothesis for an interaction of CO_2 and water relations mediated by plant hydraulics. Philos. Trans. R. Soc. B Biol. Sci. 367:583–600.

Osborne CP, Salomaa A, Kluyver TA, Visser V, Kellogg EA, Morrone O, Vorontsova MS, Clayton WD, Simpson DA. 2014. A global database of C_4 photosynthesis in grasses. New Phytol. 204:441–446.

Otto E, Young JE, Maroni G. 1986. Structure and expression of a tandem duplication of the *Drosophila* metallothionein gene. Proc. Natl. Acad. Sci. 83:6025–6029.

Otto SP. 2007. The Evolutionary Consequences of Polyploidy. Cell. 131:452–462.

Ourisson G, Nakatani Y. 1994. The terpenoid theory of the origin of cellular life: the evolution of terpenoids to cholesterol. Chemical Biology 1:11–23

Pagani M, Freeman KH, Arthur MA. 1999. Late miocene atmospheric CO_2 concentrations and the expansion of C_4 grasses. Science 285:876–9.

Panganiban G, Irvine SM, Lowe C, Roehl H, Corley LS, Sherbon B, Grenier JK, Fallon JF, Kimble J, Walker M, Wray GA, Swalla BJ, Martindale MQ, Carroll SB. 1997. The origin and evolution of animal appendages. Proc. Natl. Acad. Sci. 94:5162–6.

Papadopulos AST, Kaye M, Devaux C, Hipperson H, Lighten J, Dunning LT, Hutton I, Baker WJ, Butlin RK, Savolainen V. 2014. Evaluation of genetic isolation within an island flora reveals unusually widespread local adaptation and supports sympatric speciation. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 369:20130342.

Paradis E, Claude J, Strimmer K. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20:289–290.

Pardo–Diaz C, Salazar C, Baxter SW, Merot C, Figueiredo–Ready W, Joron M, McMillan WO, Jiggins CD. 2012. Adaptive introgression across species boundaries in *Heliconius* butterflies. PLoS Genet. 8:e1002752.

Patel RK, Jain M. 2012. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One 7:e30619.

Paterson A.H, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov A.A, Wang Y, Zhang L, Carpita N.C, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob–Ur–Rahman, Ware D, Westhoff P, Mayer KFX, Messing J, Rokhsar DS. 2009. The Sorghum bicolor genome and the diversification of grasses. Nature 457:551–556.

Peisker M. 1986. Models of carbon metabolism in C_3–C_4 intermediate plants as applied to the evolution of C_4 photosynthesis. Plant, Cell & Environment 9:627–635.

Pennell MW, Harmon LJ, Uyeda JC. 2014. Is there room for punctuated equilibrium in macroevolution? Trends Ecol. Evol. 29:23–32.

Pfaffl MW. 2001. A new mathematical model for relative quantification in. Nucleic Acids Research 29:16–21.

Phillips PC. 2008. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. Nat. Rev. Genet. 9:855–867.

Piot A, Hackel J, Christin P–A, Besnard G. 2018. One third of the plastid genes evolved under positive selection in PACMAD grasses. Planta 247:255–266.

Plachetzki DC, Fong CR, Oakley TH. 2010. The evolution of phototransduction from an ancestral cyclic nucleotide gated pathway. Proceedings. Biol. Sci. 277:1963–9.

Prasad V, Strömberg CAE, Alimohammadian H, Sahni A. 2005. Dinosaur coprolites and the early evolution of grasses and grazers. Science 310:1177–80.

Prasad V, Strömberg, CAE, Leaché, A.D, Samant B, Patnaik R, Tang L, Mohabey DM, GE S, Sahni A. 2011. Late Cretaceous origin of the rice tribe provides evidence for early diversification in Poaceae. Nat. Commun. 2:480.

Prendergast HD. V, Hattersley PW, Stone NE. 1987. New structural/biochemical associations in leaf blades of C_4 grasses (Poaceae). Funct. Plant Biol. 14:403–420.

Prentice IC, Cramer W, Harrison SP, Leemans R, Monserud RA, Solomon AM. 1992. Special Paper: A Global Biome Model Based on Plant Physiol. and Dominance, Soil Properties and Climate. J. Biogeogr. 19:117.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics, 155:945–959.

Puttick MN, Thomas GH, Benton MJ. 2014. High rates of evolution preceded the origin of birds. Evolution 68:1497–1510.

Quade J, Cerling TE, Bowman JR. 1989. Development of Asian monsoon revealed by marked ecological shift during the latest Miocene in northern Pakistan. Nature 342:163–166.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

R Development Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Rainford JL, Hofreiter M, Nicholson DB, Mayhew PJ. 2014. Phylogenetic distribution of extant richness suggests metamorphosis is a key innovation driving diversification in insects. PLoS One 9:e109085.

Ramakers C, Ruijter JM, Lekanne Deprez RH, Moorman AFM. 2003. Assumption–free analysis of quantitative real–time polymerase chain reaction (PCR) data. Neuroscience Letters 339:62–66.

Rambaut A, Suchard MA, Xie W. Drummond AJ. 2013. Tracer v1.6. Available from http://tree.bio.ed.ac.uk/software/tracer/

Reams AB, Roth JR. 2015. Mechanisms of gene duplication and amplification. Cold Spring Harbor Perspectives in Biology 7.

Reeves G, Singh P, Rossberg TA, Sogbohossou D, Schranz ME, Hibberd JM. 2018. Natural variation within a species for traits underpinning C_4 photosynthesis. Plant Physiol. pp.00168.2018.

Renvoize SA. 1982a. A survey of leaf–blade anatomy in grasses I. Andropogoneae. Kew Bull. 37:315–321.

Renvoize SA. 1982b. A survey of leaf–blade anatomy in grasses II. Arundinelleae. Kew Bull. 37:489–495

Renvoize SA. 1982c. A survey of leaf–blade anatomy in grasses III. Garnotieae. Kew Bull. 37:497.

Renvoize SA. 1985. A Note on Jansenella (Gramineae). Kew Bull. 40:470.

Renvoize SA. 1986. A Survey of Leaf–Blade Anatomy in Grasses II. Arundinelleae. Kew Bull. 41:323.

Reyna–Llorens I, Burgess SJ, Reeves G, Singh P, Stevenson SR, Williams B.P, Stanley S, Hibberd JM. 2018. Ancient duons may underpin spatial patterning of gene expression in C_4 leaves. Proc. Natl. Acad. Sci. 115:201720576.

Ripley BS, Gilbert ME, Ibrahim DG, Osborne CP. 2007. Drought constraints on C_4 photosynthesis: stomatal and metabolic limitations in C_3 and C_4 subspecies of *Alloteropsis semialata*. J. Exp. Bot. 58:1351–1363.

Ripley BS, Abraham T.I, Osborne CP. 2008. Consequences of C_4 photosynthesis for the partitioning of growth: a test using C_3 and C_4 subspecies of *Alloteropsis semialata* under nitrogen–limitation. J. Exp. Bot. 59:1705–1714.

Ripley BS, Visser V, Christin P–A, Archibald S, Martin T, Osborne C. 2015. Fire ecology of C_3 and C_4 grasses depends on evolutionary history and frequency of burning but not photosynthetic type. Ecology 96:2679–2691.

Rokas A, Carroll SB. 2008. Frequent and Widespread Parallel Evolution of Protein Sequences. Mol. Biol. Evol. 25:1943–1953.

Rondeau P, Rouch C, Besnard G. 2005. NADP–malate dehydrogenase gene evolution in Andropogoneae (Poaceae): gene duplication followed by sub–functionalization. Ann. Bot. 96:1307–1314.

Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. 61:539–542.

Rosenberg NA. 2003. The shapes of neutral gene genealogies in two species: Probabilities of monophyly, paraphyly, and polyphyly in a coalescent model. Evolution 57:1465–1477.

Rosenberg NA. 2004. DISTRUCT: A program for the graphical display of population structure. Mol. Ecol. Notes, 4:137–138.

Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, Nusbaum C, Jaffe DB. 2013. Characterizing and measuring bias in sequence data. Genome Biology 14:R51.

Saarela JM, Burke S. V, Wysocki WP, Barrett MD, Clark LG, Craine JM, Peterson PM, Soreng RJ Vorontsova MS, Duvall MR. 2018. A 250 plastome phylogeny of the grass family (Poaceae): topological support under different data partitions. PeerJ 6:e4299.

Sage RF, Li M, Monson RK. 1999. The taxonomic distribution of C_4 photosynthesis. In: Sage RF, Monson RK, editors. C_4 Plant Biology. Academic Press. p. 551–581.

Sage RF. 2001. Environmental and evolutionary preconditions for the origin and diversification of the C_4 photosynthetic syndrome. Plant Biol. 3:202–213.

Sage RF. 2004. The evolution of C_4 photosynthesis. New Phytol. 161:341–370.

Sage RF, Christin P–A, Edwards EJ. 2011. The C_4 plant lineages of planet Earth. J. Exp. Bot. 62:3155–3169.

Sage RF, Sage TL, Kocacinar F. 2012. Photorespiration and the evolution of C_4 photosynthesis. Annu. Rev. Plant Biol. 63:19–47.

Sage RF, Stata M. 2015. Photosynthetic diversity meets biodiversity: the C_4 plant example. J. Plant Physiol. 172:104–119.

Sage RF. 2017. A portrait of the C_4 photosynthetic family on the 50th anniversary of its discovery: species number, evolutionary lineages, and Hall of Fame. J. Exp. Bot. 68:e11–e28.

Sage RF, Monson RK, Ehleringer JR, Adachi S, Pearcy RW. 2018. Some like it hot: the physiological ecology of C_4 plant evolution. Oecologia 1–26.

Sage RF. 2016. A portrait of the C_4 photosynthetic family on the 50th anniversary of its discovery: species number, evolutionary lineage, and Hall of Fame. J. Exp. Bot. 67:4039–4056.

Sage TL. 2014. From proto–Kranz to C_4 Kranz: building the bridge to C_4 photosynthesis. J. Exp. Bot. 65:3341–3356.

Sánchez-García M, Matheny PB. 2017. Is the switch to an ectomycorrhizal state an evolutionary key innovation in mushroom-forming fungi? A case study in the Tricholomatineae (Agaricales). Evolution 71:51–65.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain–terminating inhibitors. Proc. Natl. Acad. Sci. 74:5463–7.

Schlötterer C, Tobler R, Kofler R, Nolte V. 2014. Sequencing pools of individuals — mining genome–wide polymorphism data without big funding. Nat. Rev. Genet. 15:749–763.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 27:863–864.

Schmitt T. 2007. Molecular biogeography of Europe: Pleistocene cycles and postglacial trends. Frontiers in Zoology, 4:11.

Shao H, Burrage LC, Sinasac DS, Hill AE, Ernest SR, O'Brien W, Courtland H–W, Jepsen KJ, Kirby A, Kulbokas EJ, Daly MJ, Broman KW, Lander ES, Nadeau JH. 2008. Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis. Proc. Natl. Acad. Sci. 105:19910–4.

Shendure J, Ji H. 2008. Next–generation DNA sequencing. Nat. Biotechnol. 26:1135–1145.

Shilla U, Tiwari BK. 2015. Impact of fire and grazing on plant diversity of a grassland ecosystem of Cherrapunjee. Keanean J. Sci. 4:67–78.

Shouliang C, Phillips SM. 2006. Microstegium Nees. In: Flora of China 22:Poaceae, pp. 593–598.

Shubin N, Tabin C, Carroll S. 2009. Deep homology and the origins of evolutionary novelty. Nature 457:818–823.

Silva C, Besnard G, Piot A, Razanatsoa J, Oliveira RP, Vorontsova MS. 2016. Museomics resolve the systematics of an endangered grass lineage endemic to north–western Madagascar. Annals of Botany 119:339–351.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E. V, Zdobnov E.M. 2015. BUSCO: assessing genome assembly and annotation completeness with single–copy orthologs. Bioinformatics 31:3210–3212.

Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. 2014. Sequencing depth and coverage: key considerations in genomic analyses. Nat. Rev. Genet. 15:121–32.

Sinha NR, Kellogg EA. 1996. Parallelism and diversity in multiple origins of C_4 photosynthesis in the grass family. Am. J. Bot. 83:1458–1470.

Smith BS, Brown WV. 1973. The Kranz syndrome in the Gramineae as indicated by carbon isotopic rations. American J. Botany, 60:505–513.

Solbrig, OT. 1996. The diversity of the savanna ecosystem. In: OT Solbrig, E Medina, and JF Silva, editors. Biodiversity and Savanna Ecosystem Processes. Ecological Studies Vol. 121:pp. 1–27. Springer–Verlag, Berlin Heidelberg.

Soros CL, Dengler NG. 2001. Ontogenetic Derivation and Cell Differentiation in Photosynthetic Tissues of C_3 and C_4 Cyperaceae. Am. J. Bot. 88:992.

Sprent JI. 2007. Evolving ideas of legume evolution and diversity: a taxonomic perspective on the occurrence of nodulation. New Phytol. 174:11–25.

Spriggs EL, Christin P–A, Edwards EJ, Kohn M, Carlini A. 2014. C_4 photosynthesis promoted species diversification during the Miocene grassland expansion. PLoS One 9:e97722.

Srivastava G, Paudayal K.N, Utescher T, Mehrotra RC. 2018. Miocene vegetation shift and climate change: Evidence from the Siwalik of Nepal. Global Planet. Change 161:108–120.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post–analysis of large phylogenies. Bioinformatics 30:1312–1313.

Stapley J, Reger J, Feulner PGD, Smadja C, Galindo J, Ekblom R, Bennison C, Ball A.D, Beckerman AP, Slate J. 2010. Adaptation genomics: the next generation. Trends Ecol. Evol. 25:705–712.

Stata M, Sage TL, Rennie TD, Khoshravesh R, Sultmanis S, Khaikin Y, Ludwig M, Sage RF. 2014. Mesophyll cells of C_4 plants have fewer chloroplasts than those of closely related C_3 plants. Plant Cell Environ. 37:2587–2600.

Stebbins GL. 1975. The role of polyploid complexes in the evolution of North American grasslands. Taxon 24:91–106.

Stern DL. 2013. The genetic causes of convergent evolution. Nat Rev Genet. 14:751–764.

Still CJ, Berry JA, Collatz G.J, DeFries R.S. 2003. Global distribution of C_3 and C_4 vegetation: Carbon cycle implications. Global Biogeochem. Cycles. 17:6–1–6–14.

Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A. 2012. Navigating the tip of the genomic iceberg: Next–generation sequencing for plant systematics. Am. J. Bot. 99:349–364.

Studer AJ, Schnable JC, Weissmann S, Kolbe AR, McKain MR, Shao Y, Cousins AB, Kellogg EA, Brutnell T.P. 2016. The draft genome of the C_3 panicoid grass species Dichanthelium oligosanthes. Genome Biol. 17:223.

Taiz L, Zeiger E. 2010. Plant Physiol. Sunderland: Sinauer Associates Inc.

Tateoka T. 1958. Notes on some grasses. VIII. On leaf structure of *Arundinella* and *Garnotia*. Bot. Gaz. 120:101–109.

Tateoka T. 1958. Notes on Some Grasses. VIII. On Leaf Structure of *Arundinella* and *Garnotia*. Bot. Gaz. 120:101–109.

Tausta SL, 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 Mol. Biol. 50:635–652.

Taylor JS, Raes J. 2004. Duplication and divergence: the evolution of new genes and old ideas. Annu. Rev. Genet. 38:615–643.

Teo SM, Pawitan Y, Ku CS, Chia KS, Salim A. 2012. Statistical challenges associated with detecting copy number variations with next–generation sequencing. Bioinformatics 28:2711–2718.

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408:796–815.

Thewissen JGM, Williams EM, Roe LJ, Hussain ST. 2001. Skeletons of terrestrial cetaceans and the relationship of whales to artiodactyls. Nature 413:277–281.

Thorne JL, Kishino H, Painter IS. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15:1647–1657.

Thorpe JP. 1982. The Molecular Clock Hypothesis: Biochemical Evolution, Genetic Differentiation and Systematics. Annu. Rev. Ecol. Syst. 13:139–168.

Tomoyasu Y, Arakane Y, Kramer KJ, Denell RE. 2009. Repeated Co–options of Exoskeleton Formation during Wing–to–Elytron Evolution in Beetles. Curr. Biol. 19:2057–2065.

True JR, Carroll SB. 2002. Gene co–option in physiological and morphological evolution. Annu. Rev. Cell Dev. Biol. 18:53–80.

Türpe AM. 1970. Sobre la anatomía foliar de *Jansenella griffithiana* (C. Mueller) Bor (Poaceae: Arundinelleae). Senckenberg. Biol. 51:277–285.

Ueno O. 1995. Occurrence of distinctive cells in leaves of C_4 species in *Arthraxon* and *Microstegium* (Andropogoneae–Poaceae) and the structural and immunocytochemical characterization of these cells. Int. J. Plant Sci. 156:270–289.

van den Bergh E, Külahoglu C, Bräutigam A, Hibberd JM, Weber APM, Zhu XG, Schranz ME. 2014. Gene and genome duplications and the origin of C_4 photosynthesis: Birth of a trait in the Cleomaceae. Curr. Plant Biol. 1:2–9.

van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. 2014. Ten years of next–generation sequencing technology. Trends Genet. 30:418–26.

Vicentini A, Barber JC, Aliscioni SS, Giussani LM, Kellogg EA. 2008. The age of the grasses and clusters of origins of C_4 photosynthesis. Glob. Chang. Biol. 14:2963–2977.

Vincens A. 1989. Palaeoenvironmental evolution of the North–Tanganyika basin (Zaire, Burundi, Tanzania). Review of Palaeobotany and Palynology, 61:69–88.

Visser V, Woodward FI, Freckleton RP, Osborne CP. 2012. Environmental factors determining the phylogenetic structure of C_4 grass communities. J. Biogeogr. 39:232–246.

von Caemmerer S, Furbank RT. 2003. The C_4 pathway: an efficient CO_2 pump. Photosynth. Res. 77:191–207.

von Caemmerer S, Quick WP, Furbank RT. 2012. The development of C_4 rice: current progress and future challenges. Science 336:1671–1672.

Voznesenskaya EV, Franceschi VR, Kiirats O, Freitag H, Edwards GE. 2001. Kranz anatomy is not essential for terrestrial C_4 plant photosynthesis. Nature 414:543–546.

Walker RP, Acheson RM, Técsi LI, Leegood RC. 1997. Phosphoenolpyruvate carboxykinase in C_4 plants: its role and regulation. Aust. J. Plant Physiol. 24:459–468.

Wang L, Czedik–Eysenberg A, Mertz RA, Si Y, Tohge T, Nunes–Nesi A, Arrivault S, Dedow L.K, Bryant DW, Zhou W, Xu J, Weissmann S, Studer A, Li P, Zhang C, LaRue T, Shao Y, Ding Z, Sun Q, Patel R. V, Turgeon R, Zhu X, Provart NJ, Mockler TC, Fernie AR, Stitt M, Liu P, Brutnell TP. 2014. Comparative analyses of C_4 and C_3 photosynthesis in developing leaves of maize and rice. Nat Biotech. 32:1158–1165.

Wang L, Peterson RB, Brutnell TP. 2011. Regulatory mechanisms underlying C_4 photosynthesis. New Phytol. 190:9–20.

Wang P, Khoshravesh R, Karki S, Tapia R, Balahadia CP, Bandyopadhyay A, Quick WP, Furbank R, Sage TL, Langdale JA. 2017. Re–creation of a Key Step in the Evolutionary Switch from C_3 to C_4 Leaf Anatomy. Curr. Biol. 27:3278–3287.e6.

Wang X, Gowik U, Tang H, Bowers JE, Westhoff P, Paterson AH. 2009. Comparative genomic analysis of C_4 photosynthetic pathway evolution in grasses. Genome Biol. 10:R68.

Wang Y, Bräutigam A, Weber APM, Zhu X–G. 2014. Three distinct biochemical subtypes of C_4 photosynthesis? A modelling analysis. J. Exp. Bot. 65:3567–3578.

Washburn JD, Schnable JC, Conant GC, Brutnell TP, Shao Y, Zhang Y, Ludwig M, Davidse G, Pires JC. 2017. Genome–guided phylo–transcriptomic methods and the nuclear phylogenetic tree of the Paniceae grasses. Sci. Rep. 7:13528.

Washburn JD, Schnable JC, Davidse G, Pires JC. 2015. Phylogeny and photosynthesis of the grass tribe Paniceae. Am. J. Bot. 102:1493–1505.

Watcharamongkol T, Christin P–A, Osborne CP. 2018. C_4 photosynthesis evolved in warm climates but promoted migration to cooler ones. Ecol. Lett. 21:376–383.

Watson L, Macfarlane TD, Dallwitz MJ. 1992 onwards. The grass genera of the world: descriptions, illustrations, identification, and information retrieval; including synonyms, morphology, anatomy, physiology, phytochemistry, cytology, classification, pathogens, world and local distribution, and references. Version: 11th December 2017.

Whitney KD, Randell RA, Rieseberg LH. 2010. Adaptive introgression of abiotic tolerance traits in the sunflower Helianthus annuus. New Phytol. 187:230–239.

Williams BP, Aubry S, Hibberd JM. 2012. Molecular evolution of genes recruited into C_4 photosynthesis. Trends Plant Sci. 17:213–220.

Williams BP, Johnston IG, Covshoff S, Hibberd JM. 2013. Phenotypic landscape inference reveals multiple evolutionary paths to C_4 photosynthesis. Elife 2.

Williams JW, Post DM, Cwynar LC, Lotter AF, Levesque AJ. 2002. Rapid and widespread vegetation responses to past climate change in the North Atlantic region. Geology 30:971.

Wingler A, Walker RP, Chen Z–H, Leegood RC. 1999. Phosphoenolpyruvate carboxykinase is involved in the decarboxylation of aspartate in the bundle sheath of maize. Plant Physiol. 120:539–546.

Wistow G. 1993. Lens crystallins: gene recruitment and evolutionary dynamism. Trends Biochem. Sci. 18:301–306.

Woodward FI, Lomas MR, Kelly CK. 2004. Global climate and the distribution of plant biomes. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 359:1465–76.

Yadav SR, Chivalkar SA, Gosavi KVC. 2010. On the identity of *Jansenella griffithiana* (Poaceae) with a new species from Western Ghats, India. Rheedea 20:38–43.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–1591.

Ye C, Hill CM, Wu S, Ruan J, Ma Z. 2016. DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. Sci. Rep. 6:31900.

Yoon S, Xuan Z, Makarov V, Ye K, Sebat J. 2009. Sensitive and accurate detection of copy number variants using read depth of coverage. Genome Research 19:1586–1592.

Yoshida S, Maruyama S, Nozaki H, Shirasu K. 2010. Horizontal Gene Transfer by the Parasitic Plant *Striga hermonthica*. Science 328:1128–1128.

Yukawa T, Ogura-Tsujita Y, Shefferson RP, Yokoyama J. 2009. Mycorrhizal diversity in Apostasia (Orchidaceae) indicates the origin and evolution of orchid mycorrhiza. Am. J. Bot. 96:1997–2009.

Zachos J, Pagani H, Sloan L, Thomas E, Billups K. 2001. Trends, rhythms, and aberrations in global climate 65 Ma to present. Science 292:686–693.

Zedane L, Hong–Wa C, Murienne J, Jeziorski C, Baldwin BG, Besnard G. 2016. Museomics illuminate the history of an extinct, paleoendemic plant lineage (Hesperelaea, Oleaceae) known from an 1875 collection from Guadalupe Island, Mexico. Biol. J. Linn. Soc. 117:44–57.

Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol. Evol. 18:292–298.

Zhaxybayeva O, Doolittle WF. 2011. Lateral gene transfer. Curr. Biol. 21:R242–R246.

Zuckerkandl E, Pauling L. 1965. Evolutionary Divergence and Convergence in Proteins. Evol. Genes Proteins:97–166.

Zuloaga FO, Morrone O, Giussani LM. 2000. A cladistic analysis of the Paniceae: a preliminary approach. In: Jacobs SWL, Everett J, editors. Grasses: systematics and evolution. CSIRO. pp. 123–135.