The
University
Of
Sheffield.

Phylogenomic investigation of lateral gene transfers among grasses

Samuel Gilbert Sidney Hibdige

A thesis submitted in partial fulfilment of the

requirements for the degree of Doctor of Philosophy

The University of Sheffield

Faculty of Science

School of Biosciences

Ecology and Evolutionary Biology

May 2022

i. Quote Page

“The outgroup is rocks.”

Joseph Felsenstein

“It's the questions we can't answer that teach us the most. They teach us how to think. If you give a man an answer, all he gains is a little fact. But give him a question and he'll look for his own answers.”

Patrick Rothfuss

“Education isn't everything, for a start it isn't an elephant.”

Spike Milligan

ii. Thesis summary

Lateral gene transfer (LGT) transcends the species barriers and allows the acquisition of genetic material that adds novelty to the recipient species. This process is widely reported in prokaryotes, but its existence in eukaryotes remained controversial until recently. The extent and adaptive significance of eukaryotic LGT remains poorly explored. In this thesis, I assess the extent of LGT among grasses and whether some groups or genes are more prone to such interspecific exchanges. I first scan for LGT in the genomes of 17 grass species covering the breadth of the family, and identify LGT in 13 of them, including wild and crop species. The rate of LGT appears higher in rhizotomous species and between closely-related groups. I then examine in further details grass genes and species in which LGT had been previously documented to evaluate the factors that promote LGT. I reconstruct the evolutionary history of an important enzyme of the C_4 pathway, and show that it has been laterally transferred at least six times in distantly-related groups of C_4 grasses. Its importance for the C_4 pathway and the requirement for gene duplications before co-option likely made LGT of this gene especially beneficial. Finally, I compare the genomes of two grass species, one of which was known to have received genes from the other, to test the hypothesis that LGT happens bidirectionally. While my analyses detected multiple LGT from the known donor, very few candidates seem to have travelled in the other direction, suggesting that LGT can be unidirectional. Overall, my work revealed that LGT is rampant among grasses, but that some genes and species are more often involved in such transfers than expected by chance. These investigations should be expanded when numerous grass genomes are available to precisely quantify the rates of LGT among lineages and across the genomes.

iii. Declaration

I, Samuel Hibdige, confirm that the thesis is my own work, unless otherwise referenced in the text. I am aware of the University's Guidance on the Use of Unfair Means (www.sheffield.ac.uk/ssid/unfair-means). This work has not previously been presented for an award at this, or any other, University and has not been submitted for any other degree. Chapter 2 has been submitted as a journal article in 'New Phytologist' and is available at doi:10.1111/nph.17328.

iv. Acknowledgements

A PhD is said to be the hardest thing you ever do that no one else will care about. That in my experience has proven to be false, the amount of care and effort that other people have shown me in the creation of this thesis truly makes me reflect on whether I could have done this without them.

My first thanks goes to my supervisor Pascal Antoine, who never once wavered in his support of me, even when I didn’t believe in myself. He only ever showed me patience, kindness and more time than at points I felt I deserved. His wide knowledge, excitement and drive is nothing but contagious to everyone around him. His constant view of the wide picture kept me grounded and without him this thesis would look very different.

My second thanks goes to my secondary supervisor, Luke Dunning, who I very much enjoyed learning from and was very generous with his time. He helped me immeasurably with my first publication, lab work, dry humour and pretty much everything in between.

In fact, my thanks is extended to the whole lab group, including, but not limited to; Jill, Matheus, Lamia and Alex who shared knowledge and made an Edwardian office with no light or climate control much more joyful.

To Emily, I owe my thanks for her endless support. I thank her for showing me what a proper work ethic is, and for pushing me towards that as close as she could. Although stressed and overworked herself, she never failed to look after me in the final days of my PhD or to make me smile.

To my PhD cohort, I thank them for welcoming me with open arms even though I was 6 months late and spoke with a southern accent. I was lucky to share my time in Sheffield with such a driven and social group of people.

To my family, I thank for their endless optimism, support and always wanting me home, and in particular dad for teaching me how to properly format a document. I would not have made it to the start of the PhD without them.

Finally, to Andrew and Elis and co, who always gave me dates to look forward to, and to the Keggins, whose ridiculousness I have found a kindred spirit in.

v. Table of Contents

Chapter 1

1 General Introduction

1.1 Lateral gene transfer as a source of novelty

An organism's genome is typically derived from its parents through sexual reproduction, in a process of vertical transmission. Novel genetic material is generated during this process by mutations as the replication of genomes is imperfect (Chandrasekaran and Betran, 2008). These mutations represent the substrate of natural selection, which leads to decreases or increases in their frequency (Mousseau and Roff, 1987, Visscher, Hill and Wray, 2008). Over time, the repeated action of natural selection on novel mutations has resulted in a remarkable array of adaptations, each caused by a myriad of selective pressures throughout an organism’s evolutionary history (Shi, Kichaev and Pasaniuc, 2016; Boyle, Li and Pritchard, 2017). While some adaptations can evolve relatively easily, the evolution of some traits requires pre-existing genes, or capacitating mutations (Blount *et al.*, 2012; Ellison and Gotelli, 2009; Schwab, 2017). In addition, the evolutionary accessibility of new traits depends on population processes, including the mutation rate, effective population sizes, and migration.

The transfer of genetic material by means other than sexual reproduction potentially allows some of the limitations of evolution through vertical descent to be bypassed. Lateral gene transfer (LGT), also known as horizontal gene transfer (HGT), is the transfer of genetic material among organisms by means other than sexual reproduction (Soucy *et al*., 2015). While LGT can theoretically occur among close relatives, the cases that receive more attention generally concern gene transfer among distantly-related species (Bergthorsson *et al*., 2004; Christin *et al*., 2012a; Husnik *et al*., 2013; Li *et al*., 2014; Lindow, 2017; Dunning *et al*. 2019). Lateral gene transfer can boost genetic diversity and increase the number of genetic variants available for selection (Sieber *et al*., 2017). LGT might therefore allow organisms to move beyond their inherent capabilities (Lindow, 2017).

The ability of LGT to act as a driving force in bacterial evolution is highlighted by the rapid spread of antibiotic genes (Sun et al., 2019). Antibiotics were first used in the 1930s but by the 1950s multidrug resistant strains of bacteria were already being reported (Davies, 1995). The emergence of such strains was far faster than would be expected based on the *de novo* rate of mutations (Ochman *et al.*, 2000) and by the 1960s it was shown that bacteria are able to transfer antibiotic resistance by LGT (Davies, 1995).

1.2 Lateral gene transfer in prokaryotes

The first experiment that alluded to the lateral transfer of genetic material between bacteria was performed in 1928 (Griffith, 1928). It was shown that a non-virulent strain of *Streptococcus pneumoniae* could become virulent when mixed with the heat-killed remains of a virulent strain. This suggested that the bacteria were able to transform into the lethal strain by using some part of the dead bacterium. These findings were followed by research in the 1930s and 40s that identified DNA as the material causing this transformation principle (Lederberg & Tatum., 1946).

In the 90 years after these experiments, many keystone discoveries in the importance of LGT in microorganism's evolution have been made. These range from the first documentation of inter-bacterial gene transfer resulting in antibiotics resistance (Ochman *et al*., 2000; Wadsworth *et al*., 2018), to the concept of pan-genomics whereby only a portion of a prokaryote genome is considered core, the rest being variable and specific to single strains (Medini *et al*., 2005; Vernikos *et al*., 2015).

There are several mechanisms for LGT that have been identified in prokaryotes. The most widely recognised being conjugation, transduction and transformation (Figure 1.1– a, c, e respectively). Conjugation is the transfer of genetic material via a structural bridge. This can only occur when there is physical contact between the donor and recipient. A single strand of DNA is transferred to the recipient's cell and subsequently used to synthesise the complementary strand, to produce a double stranded circular plasmid. Cell fusion (Figure 1.1– b) represents an advanced case of cell-to-cell contact, in which cells form aggregates that are physically joined by a fused cell membrane. During cell fusion, bidirectional gene transfer occurs that is more akin to eukaryotic sexual reproduction than prokaryotic conjugation. Transformation (Figure 1.1– e) is the uptake of exogenous DNA found within the environment. Its name derives from the transformations Griffith observed in 1928. This phenomenon has since also been observed in archaea (Worrell, Nagle, McCarthy and Eisenbraun, 1988). Transduction (Figure 1.1– c) is the transfer of DNA into a cell by means of a virus or viral vector and, as a result, does not require cell to cell contact. Unlike transformation, this mechanism protects the DNA from degradation from external DNAses. Transduction is usually a method for a virus to hijack the transcription and translation machinery of the bacterial host.

Besides these widespread mechanisms, LGT in prokaryotes can occur via gene transfer agents (GTAs). While relatively poorly understood, GTAs transfer small random pieces of genomic DNA in capsids for delivery to nearby hosts. Unlike viral transduction, GTAs are integrated into the host's chromosome, and are sometimes under regulatory control of the host. Multiple studies have shown transfer of antibiotic resistant genes across phyla using GTAs (McDaniel *et al*., 2010; Lang *et al*., 2012). However, not all bacteria, including those that can encode GTAs, are able to receive these genetic donations (Lang *et al*., 2012). It is widely presumed that GTAs have evolved from phages that have lost their ability to target their own genetic material for transfer.

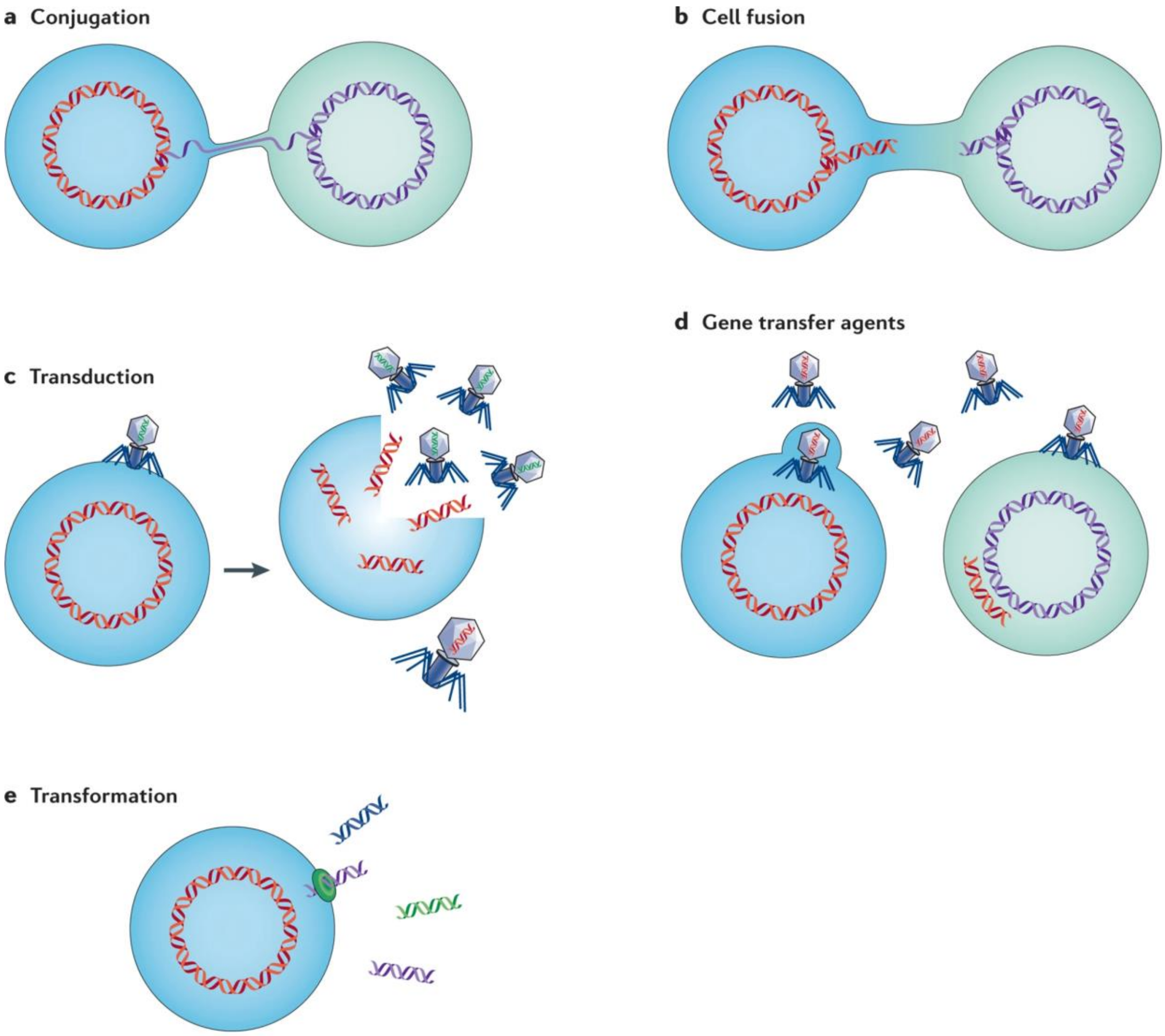


Figure 1.1: Overview of bacterial LGT mechanisms. Each panel represents a method of gene transfer. Conjugation (a) occurs through cell-cell contact whereby single stranded DNA crosses a pilus from donor to recipient. Cell fusion (b) similarly requires cell-cell contact but the transfer is bi-directional, Transduction (c) is mediated by a phage where DNA is loaded into the head. Gene transfer agents (d) are coded for by the cell's genome and transfer random pieces of genomic DNA in a manner similar to transduction. Transformation (e) is the uptake of exogenous DNA. (Reproduced from Soucy, Huang and Gogarten, 2015)

Until recently, it was assumed that LGT could occur only among closely-related prokaryotes, which have compatible systems for conjugation, higher success rate for homologous recombination, and similar codon uses (Ochman *et al*., 2000; Beiko *et al*., 2005). The transformation originally described in *S. pneumoniae* is now known to occur in over 80 species of bacteria, and transduction is now a common tool used by molecular biologists to transfer foreign DNA into a host's cell. These pieces of evidence show that capacities to accept LGT are widespread among bacteria, and reports of LGT among distantly-related bacteria have accumulated in recent years (Doolittle, 1999; Nakamura *et al*., 2004; Lerat *et al*., 2005; Cordero and Hogeweg, 2009; Wadsworth *et al.,* 2018). The width and breadth of bacterial LGT therefore shows the importance of this phenomenon for the evolution of bacteria.

1.3 Lateral gene transfer in eukaryotes

LGT in eukaryotes remains a contentious issue. As increasing numbers of complete genomes are being published, reports of LGT in eukaryotes continue to accumulate (Bergthorsson et al., 2004; Christin et al., 2012a; Husnik et al., 2013; Li et al., 2014; Bowman et al., 2017; Lindow, 2017; Dunning et al. 2019; Yang et al., 2019; Yoshida et al., 2019; Sun et al., 2020; Wang et al., 2020a; Wang et al., 2020b; Zhang et al., 2020; Zhang et al., 2020b; Cai et al., 2021; Chen et al., 2021; Mahelka et al., 2021; Park et al., 2021a Xia et al., 2021; Ma et al., 2022; Wu et al., 2022) challenging the notion that LGT only occurs between closely related or single cell organisms. The presence of a nucleus in eukaryotes makes gene exchange between mature individuals more complex as genetic material has to pass through both the cell membrane and the nuclear membrane. Furthermore, any DNA insertion will not be passed onto subsequent generations, unless it is integrated within the germline or undifferentiated cells capable of vegetative propagation. However, getting into these cells can be challenging if the organism has specialised germline tissue as in vertebrates.

The claims of LGT however often come with controversy due to potential technical problems, such as contamination and analytical errors (Danchin, 2016; Martin, 2017). However, recent studies that have ruled out such biases still identified a number of LGT in eukaryotes (Christin et al., 2012a; Lindow, 2017; Dunning et al., 2019). Moreover, there is published evidence that LGT among eukaryotes can add functional diversity to the recipient genome (Bergthorsson et al., 2004; Christin et al., 2012a; Husnik et al., 2013; Li et al., 2014; Lindow, 2017; Dunning et al. 2019; Xia et al., 2021; Wu et al., 2022). The role of LGT in eukaryotic evolution remains poorly explored, yet this process provides unique opportunities to assess how novel, major mutations represented by gene transfers can affect evolutionary trajectories.

The first reports of LGT among eukaryotes were thought to be special cases associated with intimate interspecies relationships, usually among parasites and their hosts. The most widespread instances of prokaryotic-to-eukaryotic LGT concerns obligate endosymbionts, such as the eukaryotic organelles derived from α-proteobacteria (mitochondria) and cyanobacteria (chloroplast) (Boucher *et al*., 2003).

Obligate endosymbionts have transferred genes directly into the nuclear genome of the hosts (Timmis *et al*., 2004). The organelle DNA exists in remission, slowly relocating to the nucleus (Timmis *et al*., 2004; Kleine, Maier and Leister, 2009). The variation in the amount of genes remaining within the organelles across the plant kingdom shows that gene transfer to the nucleus is an ongoing process (Cullis *et al*., 2008). Plant chloroplasts for example, only retain 60 -100 genes out of the ~1,500 existing in free living cyanobacteria, and it is estimated that between 11 and 14% of the nuclear DNA of *Cyanophora* and *Arabidopsis* has been acquired from the chloroplasts (Reyes-Prieto *et al*. 2006; Deusch *et al*. 2008; Nowack *et al*., 2010).

Examples of prokaryotic-to-eukaryotic LGT however are not limited to organelles or endosymbionts, but can span vast evolutionary distances. Analysis of the sweet potato genome has shown that all analysed accessions contained one or more transfer DNA (T-DNA) sequences, tumour inducing plasmids, originating from *Agrobacterium tumefaciens*, (Kyndt *et al*., 2015). Not only were the sequences expressed, these insertions were not present in the nearest wild relatives, suggesting that the LGT were selected for during domestication. However, the traits they are associated with have not been identified (Kyndt *et al*., 2015). *Agrobacterium* spp. are already used experimentally to transform plants, but the fact that the plant has used natural T-DNA inserts to its advantage is novel (Kyndt *et al*., 2015).

More often than not, examples of functional prokaryotic-to-eukaryotic LGT generally concern the transfer of a single gene or pathway from a single donor. However, in some cases, pathways of eukaryotes were assembled from multiple LGT, for example in the mealybug *Planococcus citri* (Husnik *et al*., 2013). At least 22 laterally acquired genes exist from multiple diverse bacteria, none of which originate from an obligate symbiont of *P. citri*. (Husnik *et al*., 2013). It has even been suggested that major episodes of horizontal gene transfer drove the evolution of land plants (Ma *et al*., 2022).

Eukaryote-to-eukaryote LGT are rarer in the literature and historically reported as chance discovery. Parasitic eukaryote LGT were thought to be an anomaly where host and parasite LGT was made possible due to prolonged physical association with each other. For the parasite, it is hypothesised that transcription of host genes may aid resource extraction and reduce the effectiveness of a host's response and ability to mount a defence (de Felipe et al., 2005). There are several examples of eukaryote host-to-parasite LGT, including in *Rafflesia cantleyi* (Xi *et al*., 2012) and *Striga hermonthica*, a pervasive crop parasite (Yoshida *et al*., 2010, 2019). Gene flow does also occur in the opposite direction whereby hosts have received genes from parasites (Mower *et al*., 2004; Davis, Anderson and Wurdack, 2005), but the benefits of these LGTs are not generally understood.

Parasitism is not a prerequisite of eukaryote-to-eukaryote LGT, but the mechanism of transfer remains elusive in non-parasitic groups. In animals and fungi for instance, the pea aphid *Acyrthosiphon pisum* has acquired the genes for carotenoid biosynthesis that most animals lack, from a fungal genome (Nováková and Moran, 2011). This gives the aphid its characteristic orange colouring and is thought to help in camouflage against visual predation. The sweet potato whitefly *Bemisia tabaci* has been shown to have laterally acquired a plant detoxification gene, which allows it to neutralise plant defence compounds enabling a wide polyphagous diet (Xia *et al*., 2021). Retrotransposons have also been shared between bivalves and other aquatic species of multiple phyla (Metzger *et al*., 2018). In fungi, three hallucinogenic mushroom genomes contain a shared psilocybin gene cluster that provides evidence for LGT between lineages (Reynolds *et al*., 2018).

Non-parasitic plants also contain many examples of LGT. Entire mitochondrial genes of three different green algae species and one moss have been identified in the mitochondrial genome of the angiosperm *Amborella trichopoda* (Bergthorsson *et al.*, 2004; Lindow, 2017). Examples of eukaryote-to-eukaryote LGT of nuclear genes have also been reported. For instance, ferns possess a neochrome photoreceptor from hornworts, postulated to help cope with the low light conditions caused by the appearance of an angiosperm canopy (Li *et al*., 2014). In addition, several of the key enzymes involved in the C_4 photosynthetic pathway of the grass *Alloteropsis semialata* appear to have been laterally acquired from other distantly related grasses (Figure 1.2; Figure 1.3; Christin *et al*., 2012a; Dunning *et al*., 2019).

There is growing evidence that eukaryote-to-eukaryote LGT is more frequent than originally thought, and in some cases it can have adaptive consequences. The extent of this phenomenon remains, however, poorly understood, as previous cases were typically incidentally identified, and dedicated efforts to identify LGT and their consequences are largely lacking.

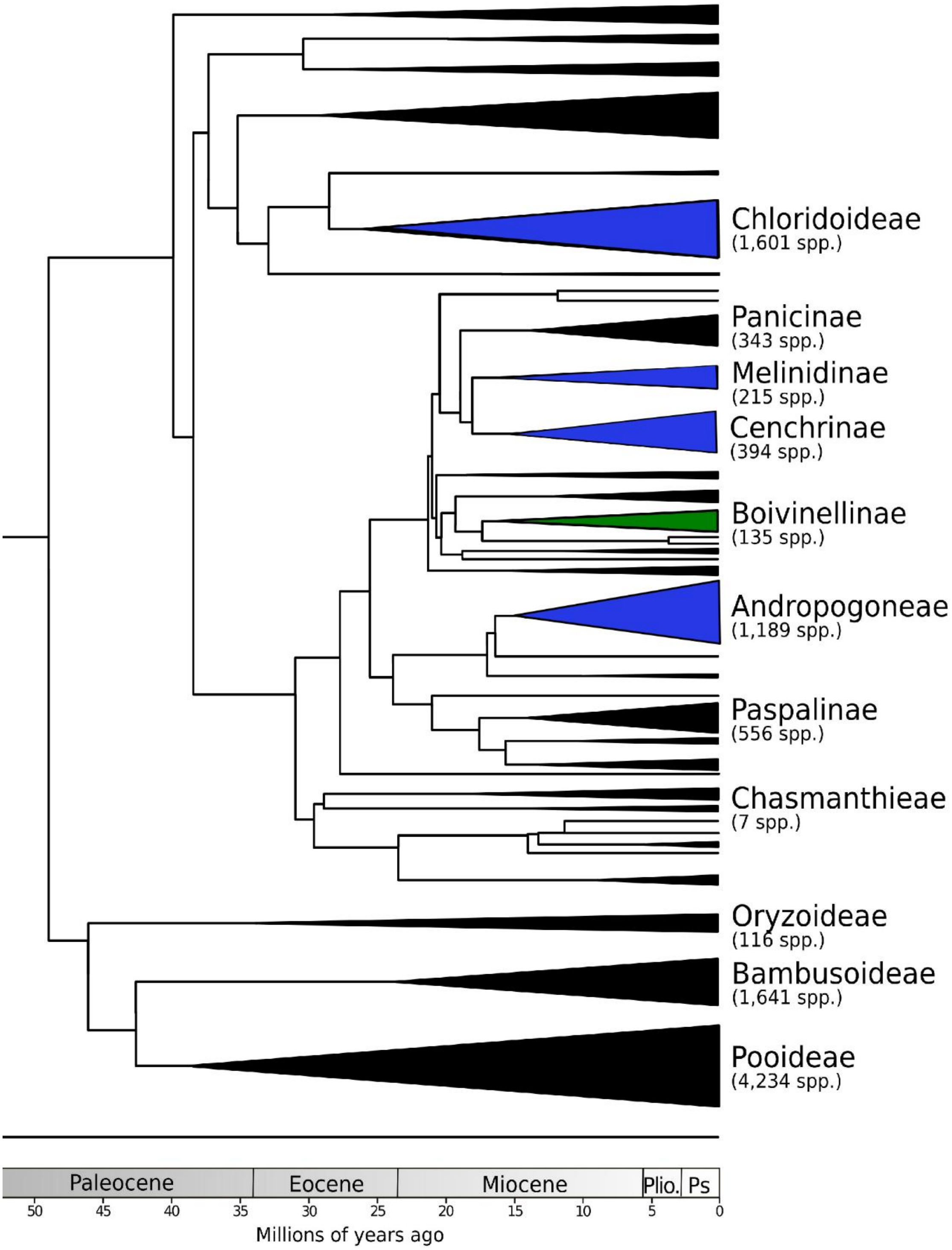


Figure 1.2: Grass phylogeny of key grass groups, compiled from Christin *et al.*, 2013. Species numbers come from Soreng *et al.*, 2015. *Alloteropsis* is located within Boivinellinae (green) and has received LGT from other grass groups (blue) (Christen *et al.*, 2012; Dunning *et al.*, 2019)

1.4 Grasses as a study system

While LGT has been detected in various groups of eukaryotes, few multicellular systems are better than grasses to study this phenomenon. The grass family (Poaceae) contains over 12,000 species (Soreng *et al*., 2015) exhibiting a diverse range of phenotypes that may contribute to LGT dynamics. The grass family as a group is of particular interest to humans as it contains many agricultural crops that are important global food sources, including rice, barely, wheat, maize, sorghum and millet. In fact, only three grasses (rice, wheat and maize) contribute more than half of the world's calorie intake (Allender, 2011). Besides the importance as a primary source of human food, grasses constitute whole or partial diets for domestic animals as fodder (O'Mara, 2012; Fuglie, Peters, and Burkart, 2021). Grasses may also act as a potential source for biofuel production, in particular C_4 grasses are of specific interest due to their high productivity and resource use efficiency (van der Weijde et al., 2013). As a result, there has already been extensive research into grass genetics, evolution and biochemistry, and multiple full genomes are available for this family.

As stated previously, a good example of functional LGT was reported in a grass belonging to the genus *Alloteropsis,* including *Alloteropsis semialata*, a perennial grass disturbed across much of tropical and subtropical Africa, Asia and Australia (Figure 1.3). An initial study based on Sanger and 454 sequencing showed that two key enzymes of the C_4 pathway of *Alloteropsis* had been laterally acquired from two distant grass genera (Christin *et al*. 2012). Further studies showed that these LGT happened during the diversification of *Alloteropsis*, with subsequent introgression among species (Olofsson *et al*. 2016; Dunning *et al*. 2017). These LGT are thought to have been adaptive as they allowed shortcutting of the evolution of C_4 enzymes via natural selection (Phansopa *et al*., 2020). Indeed, the complex C_4 pathway results from the co-option of multiple enzymes followed by the adaptation of their kinetic properties via adaptive amino acid changes (Christin *et al*. 2007; Huang *et al*. 2017). Because the genes laterally acquired came from grass lineages that had evolved the C_4 trait millions of years before, they were already adapted for the C_4 function, therefore providing a fitness advantage for *Alloteropsis* (Christin *et al*. 2012). This initial report was followed by a large genome-wide analysis of *Alloteropsis semialata* which identified another 57 genes laterally acquired from at least nine different grasses (Dunning *et al*. 2019). This previous effort also incidentally obtained evidence for LGT among grasses other than *Alloteropsis* (Dunning *et al*. 2019). These findings add to evidence of LGT reported for other grasses, including the transfer of ribosomal DNA into *Hordeum* species (Mahelka *et al*., 2017), and possible LGT among other grasses (Vallenback *et al*. 2010; Park, Christin and Bennetzen, 2021a; Mahelka et al., 2021; Wu et al., 2022).

Previous reports of LGT among grasses were mostly discovered incidentally, but LGT have still been reported for multiple lineages. These results suggest that the known cases might just be the tip of the iceberg. Grasses originated at least 50 million years ago (Christin *et al.*, 2014), and include more than 12,000 species spread on all continents. Together, grasses cover 20% of the land's surface (Shantz, 1954). Many grasses exhibit vegetative reproduction, and LGT in tissues involved in such growth will be maintained in next year's growth, effectively becoming part of the germline. Moreover, grasses are wind pollinated, providing numerous opportunities for cell-to-cell contact following illegitimate pollination. Finally, grasses are famous for their high content of transposable elements and their dynamic genomes that undergo frequent rearrangements (Schnable *et al.*, 2009; Park *et al.*, 2011b; Park *et al*., 2011a; Kim *et al*., 2014; Park *et al*., 2021) . All these properties might facilitate the transfer of DNA and its subsequent integration into the nuclear genomes.

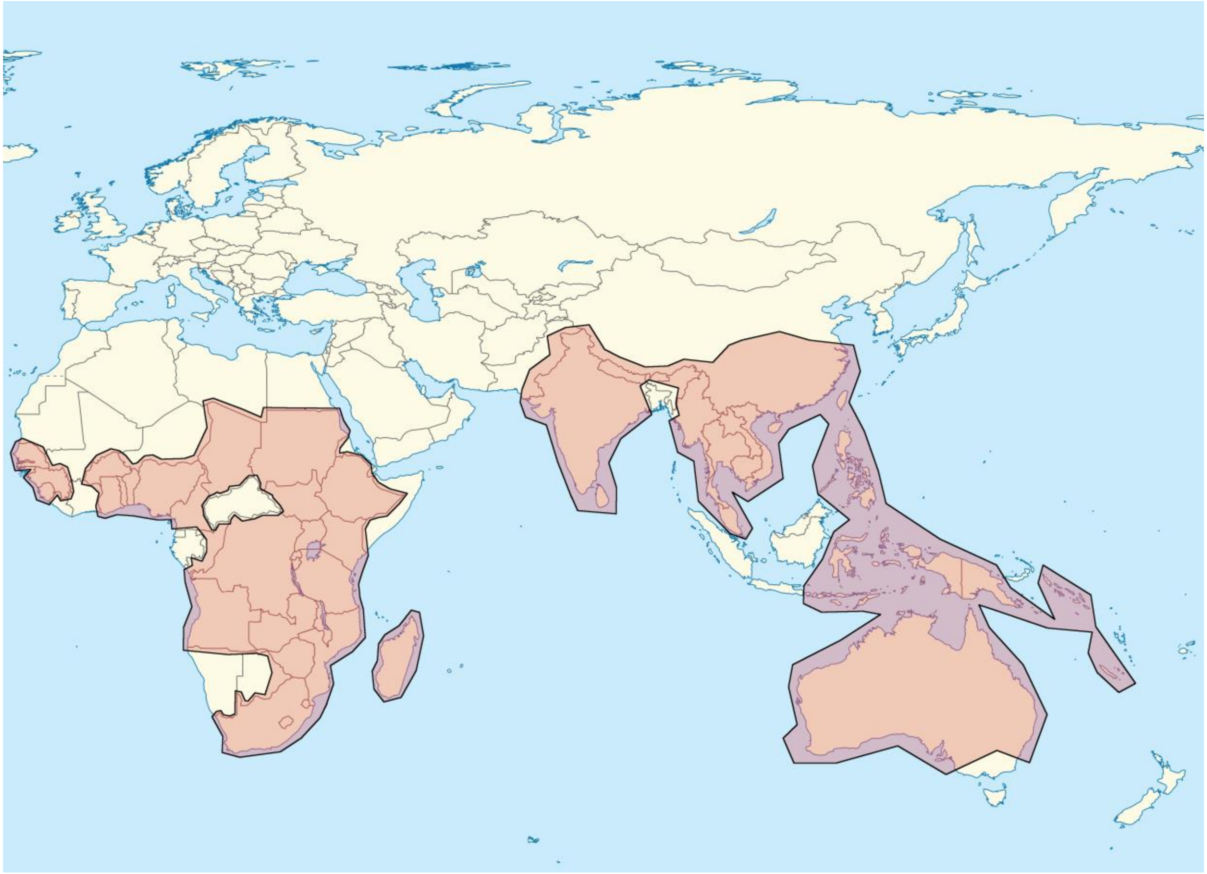


Figure 1.3: The natural range of *Alloteropsis semialata*, The range data is to country level (*Alloteropsis semialata* (R.Br.) Hitchc. | Plants of the World Online | Kew Science, 2022) and overlaid on general map data (Natural Earth - Free vector and raster map data at 1:10m, 1:50m, and 1:110m scales, 2022)

1.5 Methods to detect LGT

Lateral gene transfers are generally identified as genes in a given organism that are more similar to those of distant related species than to those of close relatives. Evaluating sequence similarities can be done directly, for example using alignment-free approaches, which have been developed in recent years for prokaryotes. For instance, BLAST-related methods such as the ALFY method (Alignment-Free Local homologY) use unique identified substrings of DNA due to the difficulties in creating bacterial alignments (Domazet-Lošo and Haubold, 2011). Other text-based mining methods can be used that detect unusual regions within a string without any domain knowledge, and there have been attempts to use this to detect LGT (Taniguchi *et al.*, 2013). These alignment-free methods have yet to be tested on eukaryotes, and only consider one target sequence without taking into account any group structure (taxonomic or ecological structure; Cong, Chan and Ragan, 2016).

While similarity analyses allow rapid scans of genomes, unequivocal evidence for LGT is typically provided by phylogenetic trees. Indeed, LGT would create a strong conflict between species and gene trees, as reported in several instances (Christin *et al*. 2012; Dunning *et al*., 2019). Species trees can be inferred from different sets of markers that are sufficiently conserved to be compared across large taxonomic groups. While such species trees historically relied on a few markers due to sequencing difficulties, the advent of high-throughput sequencing has provided large genomic datasets, so that species trees can now be inferred from a wider number of species and larger number of genes (One Thousand Plant Transcriptomes Initiative, 2019; Williams et al., 2019; Kawahara et al., 2019)

Using phylogenetics, lateral gene transfers would be evidenced by conflicts between species and gene trees, but other processes can give rise to the same pattern (Pamilo & Nei,. 1988; Felsenstein, 1988; Nichols, 2001). First, incomplete lineage sorting results from the maintenance of ancestral polymorphisms across speciation events (Figure 1.4, C; Degnan & Rosenberg, 2009). It is estimated that up to 30% of the human genome is more closely related to gorillas than chimpanzees due to this process (Scally *et al*. 2012). Second, gene duplication followed by losses can also cause discordant gene trees if different paralogs are lost in each of the clades, a classical paralogy problem (Figure 1.4, B; Szöllősi and Daubin 2012). Third, systematic biases can lead to erroneous gene trees. Convergent evolution due to adaptation or mutation biases can cause genes within distantly related species to appear deceptively closely related, as demonstrated by convergent feeding adaptations in red pandas and giant pandas (Hu *et al*., 2017) or genes for C_4 photosynthesis (Christin *et al*,. 2007). Fourth, contamination continues to be a big problem in LGT detection. The most prolific example is the initial publication of the tardigrade genome, which claimed that 17% of its genes originated from LGT (Boothby *et al*., 2015). A later reanalysis brought this number to 0.4%, showing that the difference was due to contamination problems (Koutosvoulos *et al*., 2016). Finally, poor alignments and erroneous sequences can contribute to wrong gene trees. Ruling out these potential problems requires extra phylogenetic considerations.

Incomplete lineage sorting can be ruled out by considering LGT among distantly-related species. Indeed, the importance of incomplete lineage sorting depends on the pace of speciation events, and this phenomenon is unlikely to concern species separated by large evolutionary distances (Maddison, 1997). Paralogy problems require a good assessment of orthology, which is better achieved with a dense species sampling coupled with genomic information. In addition, molecular dating can be used to show that the gene divergence is more recent than the species divergence, ruling out paralogy problems as an alternative explanation (Christin *et al*. 2012c). Ruling out systematic biases can be achieved by comparing different data partitions. For instance, 3rd positions of codons and introns are less subject to selection, and can therefore be used to exclude convergent adaptive evolution (Bofkin and Goldman, 2006). The risk of contamination should be minimised by repeating the sequencing efforts on different samples, and if possible, in different labs (Christin *et al*, 2012). Because biological samples of multicellular species generally contain microbial organisms, ruling out contamination is extremely challenging when studying prokaryotic-to-eukaryotic LGT. This problem is strongly reduced by focusing on plant-to-plant LGT, as repeated contamination is unlikely. Finally, alignment and sequence errors are better considered by carefully inspecting the alignments and trees. Overall, a robust phylogenomic approach is required to confirm that LGT is indeed being observed.

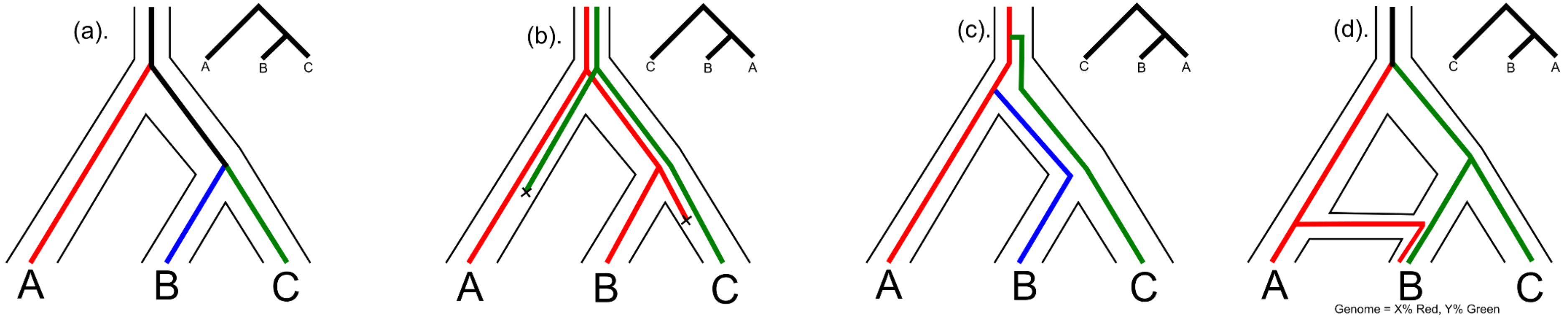


Figure 1.4: Sources of gene tree–species tree discordance. The small phylogeny in each window represents the observed gene tree. (a) No discordance, gene tree matches species tree. (b) Gene duplication and loss: through extinction of lineages, gene duplication can produce apparent relationships incongruent with the species tree. Even if paralogs are not lost, the sampling of lineages that are not true orthologs can cause A and B to appear more closely related to each other than either is to C. (c) Incomplete lineage sorting, due to loss of alleles after speciation, A and B will seem more closely related as their orthologs diverged more recently than the one present in C resulting in the species tree ((A, B), C). (d) Hybridization causes some genes sampled from species B to descend from the population ancestral to A, whereas others descend from the population ancestral to B and C. Gene trees will therefore depict either ((A, B),C) (red) or (A,(B, C)) (green) depending on which parent the gene originated from. Hybridization, at first, affects the whole genome. After multiple backcrosses, it can result in only few genes from one lineage remaining and might therefore look like LGT. Modified from Degnan & Rosenberg, 2009.

1.6 Thesis aims, objectives and outline

In this thesis, I use various approaches to investigate the frequency of LGT among different groups of grasses to assess their potential evolutionary significance. I capitalise on available genomic resources, and combine whole-genome scans with detailed analyses of genes or species that have previously been shown to be involved in LGT. The available genomic resources are completed by new data generation, where needed. My research is divided in three data chapters, which address interrelated questions and together give new insights into the frequency of grass-to-grass LGT and the factors that might make some genes or species more prone to interspecific gene exchanges.

In Chapter 2, I scan the genomes of 17 grass species that span more than 50 myrs of evolution for LGT. These species include major crops as well as wild grasses, and represent different ploidy levels, growth strategies and geographic origins. Using phylogenetic approaches, I first identify LGT received by each of the 17 species. I then use comparative approaches to test for an effect of different species properties on the amount of LGT. Finally, I compare the amount of LGT among groups that represent different phylogenetic distances. These analyses reveal that LGT is widespread in the family and point to the importance of phylogenetic relatedness, shared geographic ranges, and rhizomatous growth in promoting LGT in the group.

In Chapter 3, I focused my attention on one gene that has previously been shown to have been laterally-transferred among grasses; the gene for the C_4 photosynthetic enzyme phosphoenolpyruvate carboxykinase (PCK). I reconstruct the history of PCK genes using the genome and transcriptome data available for a large number of grass species. Comparison of the PCK gene tree with the expected species relationships identifies multiple cases of potential LGT. I then use molecular dating to confirm the LGT scenario. These analyses show that the same gene has been repeatedly transferred among several groups of grasses, and I discuss the factors that might have made this gene especially prone to LGT.

In Chapter 4, I develop a novel similarity-based approach to quickly compare the genomes of two grasses and detect potential LGT. One of these two species has previously been shown to have received LGT from the other, and my analyses were especially designed to (1) detect non-coding LGT and (2) test the hypothesis that LGT happens bidirectionally among pairs of species. Using a *de novo* genome I generated for the second species, I find that LGT happened mainly in one direction among these two species, showing that the factors making some species prone to receive LGT are different from those that make some species prone to give LGT.

These three data chapters show that LGT is rampant within the grass family, with some species and some genes especially likely to be involved in interspecific transfers. I discuss the caveats and power of the analyses and their joint significance in the general discussion provided at the end of the thesis.

Chapter 2

2 Widespread lateral gene transfer among grasses

Edited version of the submitted manuscript as seen in New Phytologist (2021). Volume 230, issue 6. Pages 2474-2486.

Samuel G. S. Hibdige[1], Pauline Raimondeau[1], Pascal-Antoine Christin[1], Luke T. Dunning[1]

Keywords: adaptation, evolution, genomics, horizontal gene transfer, phylogenomics, Poaceae.

Affiliations: [1] Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom

Personal contributions: I performed the analyses of the samples belonging to the Paniceae tribe, while Luke Dunning analysed the other samples. I wrote the manuscript with Luke Dunning.

2.1 Summary

Lateral gene transfer (LGT) occurs in a broad range of prokaryotes and eukaryotes, in some cases promoting adaptation. LGT of functional nuclear genes has been reported among some plants, but systematic studies are needed to assess the frequency and facilitators of LGT. We scan the genomes of a diverse set of 17 grass species that span more than 50 million years of divergence and include major crops to identify grass-to-grass protein-coding LGT. We identify LGTs in 13 species, with significant variation in the amount each received. Rhizomatous species acquired statistically more genes, probably because this growth habit boosts opportunities for transfer into the germline. In addition, the amount of LGT increases with phylogenetic relatedness, which might reflect genomic compatibility amongst close relatives facilitating successful transfers. However, genetic exchanges among highly divergent species indicate that transfers across almost the entire family can occur. Overall, we show that LGT is a widespread phenomenon in grasses, which has moved functional genes across the grass family into domesticated and wild species alike. Successful LGTs appear to increase with both opportunity and compatibility.

2.2 Introduction

The adaptive potential of a species is limited by its evolutionary history, the amount of standing genetic variation and the rate of new mutations (Barrett & Schluter, 2008). Lateral gene transfer (LGT) enables organisms to overcome these limitations by exchanging genetic material between lineages that have evolved significant reproductive barriers (Doolittle, 1999). LGT is an important evolutionary force in prokaryotes, with up to 60% of genes within a species pan-genome acquired in this manner (Freschi *et al.*, 2018). The genes transferred can have a dramatic effect on adaptation, facilitating the colonisation of new niches and the development of novel phenotypes, as exemplified by the rapid spread of antibiotic resistance in bacteria (Ochman *et al.*, 2000). While LGT is more prevalent in prokaryotes, it has also been documented in a variety of multicellular eukaryotes (reviewed in: Anderson, 2005; Keeling & Palmer, 2008; Schönknecht *et al.*, 2014; Husnik *et al.*, 2018; Van Etten & Bhattacharya, 2020), including plants (reviewed in: Richardson & Palmer, 2007; Gao *et al.*, 2014; Wickell & Li, 2019; Chen *et al.*, 2021).

DNA has been transferred into plants from prokaryotes, fungi and viruses, in particular with recipients in algae (Cheng *et al.*, 2019) and bryophytes (Yue *et al.*, 2012; Maumus *et al.*, 2014; Bowman *et al.*, 2017; Zhang *et al.*, 2020). Concerning plant-to-plant transfers, a majority of nuclear LGTs reported so far involve the transfer of genetic material between parasitic species and their hosts, with examples from the genera *Cuscuta* (Vogel *et al.*, 2018; Yang *et al.*, 2019), *Rafflesia* (Xi *et al.*, 2012), and *Striga* (Yoshida *et al.*, 2010, 2019). However, plant-to-plant LGT is not restricted to parasitic interactions, and it has been recorded in ferns (Li *et al.*, 2014) and eight different species of grasses (Vallenback *et al.*, 2008; Christin *et al.*, 2012a; Prentice *et al.*, 2015; Mahelka *et al.*, 2017; Dunning *et al.*, 2019). Grasses represent one of the best systems to investigate factors promoting LGT between non-parasitic plants as multiple transfers have been identified in the group, and there is extensive genomic resources available due to their economic and ecological importance (Chen *et al.*, 2018). Early examples of grass-to-grass LGT were largely obtained incidentally, and only one grass genome (*Alloteropsis semialata*) has been comprehensively scanned, with 59 LGTs identified using stringent phylogenetic filters (Dunning *et al.*, 2019). These 59 protein-coding genes were transferred from at least nine different donors as part of 23 large fragments of foreign DNA (up to 170 kb per fragment). A majority of the acquired LGTs within *A. semialata* are expressed, with functions associated with photosynthesis, disease resistance and abiotic stress tolerance (Dunning *et al.*, 2019; Phansopa *et al.*, 2020). While reports of LGT in other species in the group suggest it is a widespread phenomenon, its full distribution within the family remains to be assessed.

Grasses are very diverse (Soreng *et al.*, 2015), with more than 12,000 species exhibiting extensive phenotypic variation that may influence LGT dynamics. In particular, the family contains both annuals and perennials. If LGT happens during vegetative growth e.g. root-to-root inosculation (Dunning *et al.*, 2019), or other graft-like processes (Stegemann and Bock, 2009; Hertle *et al.*, 2021), the number of LGTs is predicted to be higher in perennial and rhizomatous species. Conversely, if LGT happens through illegitimate pollination (Christin *et al.*, 2012a), the number of LGTs may not vary with growth form as the wind-pollinated syndrome is universal in this group, or it could be higher in annuals that produce seeds more frequently. The frequency of LGT between species is also likely to be influenced by their geographical distribution, as transfers require the physical movement of DNA. The mechanism of transfer will dictate whether the minimal distance lies within the zone of direct contact (e.g. in the case of inosculation) or within the limits of pollen dispersal (e.g. in the case of illegitimate pollination). Finally, successful transfers might be more likely to occur between closely-related groups with similar genome features as observed in prokaryotes (Skippington & Ragan, 2012; Soucy *et al.*, 2015). Most of the grass diversity is clustered in the two BOP and PACMAD sister groups that diverged more than 50 million years ago (Christin *et al.*, 2014). Each of the two groups has more than 5,000 taxa and includes model species with complete genomes (Soreng *et al.*, 2015). The family therefore offers unparalleled opportunities to assess whether functional characteristics or phylogenetic distance determines the amount of LGT among non-parasitic plants.

In this study, we use a phylogenomic approach to scan 17 different grass genomes and quantify LGT among them. The sampled species belong to five different clades of grasses, two from the BOP group (Oryzoideae and Pooideae) and three from the PACMAD group (Andropogoneae, Chloridoideae, and Paniceae). Together, these five groups contain more than 8,000 species or over 70% of the diversity within the whole family (Soreng *et al.*, 2015). In our sampling, each of these five groups is represented by at least two divergent species, allowing us to monitor the number of transfers among each pair of groups. In addition, the sampled species represent a variety of domestication statuses, life-history strategies, genome sizes, and ploidy levels (Table 2.1). Using this sampling design, we (i) test whether LGT is more common in certain phylogenetic lineages, and (ii) test whether some plant characters are associated with a statistical increase of LGT. We then focus on the donors of the LGTs received by the Paniceae tribe, a group for which seven genomes are available, to (iii) test whether the number of LGTs increases with phylogenetic relatedness. Our work represents the first systematic quantification of LGT among members of a large group of non-parasitic plants and sheds new light on the conditions that promote genetic exchanges across species boundaries in plants.

2.3 Materials and Methods

2.3.1 Detecting grass-to-grass LGT

We modified the approach previously used by Dunning *et al.*, (2019) to identify grass-to-grass LGT. Specifically, the initial mapping filtering step was discarded to avoid preferentially detecting LGTs in groups for which high-coverage genome data are available for multiple closely related species. In total, 17 genomes were scanned for LGT (Table 2.1), with all phylogenetic analyses based on coding sequences (total = 817,621 genes; mean per species 48,095 genes; SD = 26,764 genes). Our analytical pipeline relies on BLAST searches followed by phylogenetic inference and filtering based on phylogenetic patterns, which is analogous to existing tools to identify putative orthologs (Emms *et al.*, 2015). Using our custom pipeline allowed us to tailor its details to the purpose of identifying putative LGT from any type of gene family. Furthermore, we perform additional synteny analysis to verify that our method recovers true orthologs.

As a first step, we verified for each gene whether its relationships to the best-hit match from 36 other species were as expected based on the species tree, to rapidly discard genes that are clearly not LGT and focus subsequent analyses on plausible candidates. In this step, 37-taxa trees were constructed using data from the 17 grass genomes (Table 2.1), supplemented with transcriptome data for 20 additional species from across the grass family (Moreno-Villena *et al.*, 2018; Supplementary Table 2.1). For each gene, we used BLASTn to identify the best hit (highest bit-score) with a minimum match length of 300bp (not necessarily a single continuous BLAST match) from each of the other 36 species. These sequences were then extracted and nucleotide alignments were generated by aligning the BLASTn matching regions to the query sequence using the 'add fragments' parameter in MAFFT v7.427 (Katoh and Standley, 2013). If the BLASTn match for a species was fragmented, the different fragments were joined into a single sequence after they had been aligned. Alignments with less than ten species were considered non informative and consequently discarded (retained 55.9% of genes; total = 457,003 genes; mean per species 26,883 genes; SD = 13,042 genes; Supplementary Table 2.2). For each alignment with ten species or more, a maximum-likelihood phylogenetic tree was inferred using PhyML v.20120412 (Guindon and Gascuel, 2003) with the GTR+G+I substitution model. Each topology was then mid-point rooted using the phytools (Phylogenetic Tools for Comparative Biology (and Other Things)) package in R and Perl scripts (available from GitHub: https://github.com/SamuelHibdige/) were used to identify genes from each focus species nested within a different group of grasses. We focused on five groups (Andropogoneae, Chloridoideae, Oryzoideae, Paniceae and Pooideae) represented by at least two complete genomes that were supported by most gene trees in a previous multigene coalescent species tree analysis (Figure 2.1; Dunning *et al.*, 2019). The whole set of analyses were later repeated to detect LGT between well supported subclades within the Paniceae, the most densely sampled group with seven genomes spread across the group (Figure 2.1). In these subsequent analyses, we considered LGTs received from two Paniceae clades represented by two genomes and supported by most gene trees in previous analyses (i.e. Cenchrinae and Panicinae, Figure 2.1; Dunning *et al.*, 2019). To be considered as nested, the sister group of the query gene (joining at node 1), and their combined sister group (joining at node 2), had to belong to the same grass group to which the query gene does not belong. For genes that were nested, the analysis was repeated with 100 bootstrap replicates produced by PhyML to verify that the nesting of the query sequence was supported by bootstrap node support values of at least 50% at either node 1 or node 2. A soft bootstrap node support threshold (50%) was used to retain all potential LGTs for the more stringent second filtering step (see Supplementary Figure 2.1 for the impact of varying this threshold).

For candidates that passed the first phylogenetic filter, we performed a second round of filtering using data from 105 genome/transcriptome datasets belonging to 85 species, including the datasets used for the 37-taxa trees (Supplementary Table 2.1). For each LGT candidate, we used BLASTn to identify all matches (not just the best match) with a minimum alignment length of 300bp (not necessarily a single continuous blast match) in each of the 105 datasets. Alignments were generated as previously, before being re-aligned as codons using MAFFT and manually trimmed with a codon-preserving method to remove poorly aligned regions. Maximum likelihood phylogenies were then inferred using PhyML v.21031022, with the best substitution model identified by Smart Model Selection SMS v.1.8.1 (Lefort *et al.*, 2017). The trees were manually inspected and discarded if: i) there were too few taxa with either less than three species within the LGT donor clade, or less than three species outside the LGT donor clade; ii) the LGT candidate was not nested within another group of grasses with the increased taxon sampling; or iii) the tree had obvious paralogy problems due to gene duplication events. For retained candidates, we removed paralogs representing duplicates originating before the core grasses (BOP and PACMAD clades; Soreng *et al.*, 2015), and joined fragmented transcripts from a single data set if they were nested within the same phylogenetic group. To avoid merging recent paralogs we retained separate transcripts if they overlapped significantly and had multiple nucleotide substitutions. Up to this point the analyses were performed on each gene from each genome, and an individual phylogenetic tree was thus computed for each gene belonging to a group of recent duplicates (e.g. generated by allopolyploidization). For subsequent downstream analyses, we only retained one gene tree per group of homologous LGT candidates (e.g. taxon-specific duplicates). The tree inference was then repeated with 100 bootstraps, and the trees were again manually inspected, retaining candidates where the placement of the LGT in a group was supported by at least one node with ≥ 70% bootstrap node support. Finally, BLASTx was used to annotate the LGT candidates against the SwissProt database.

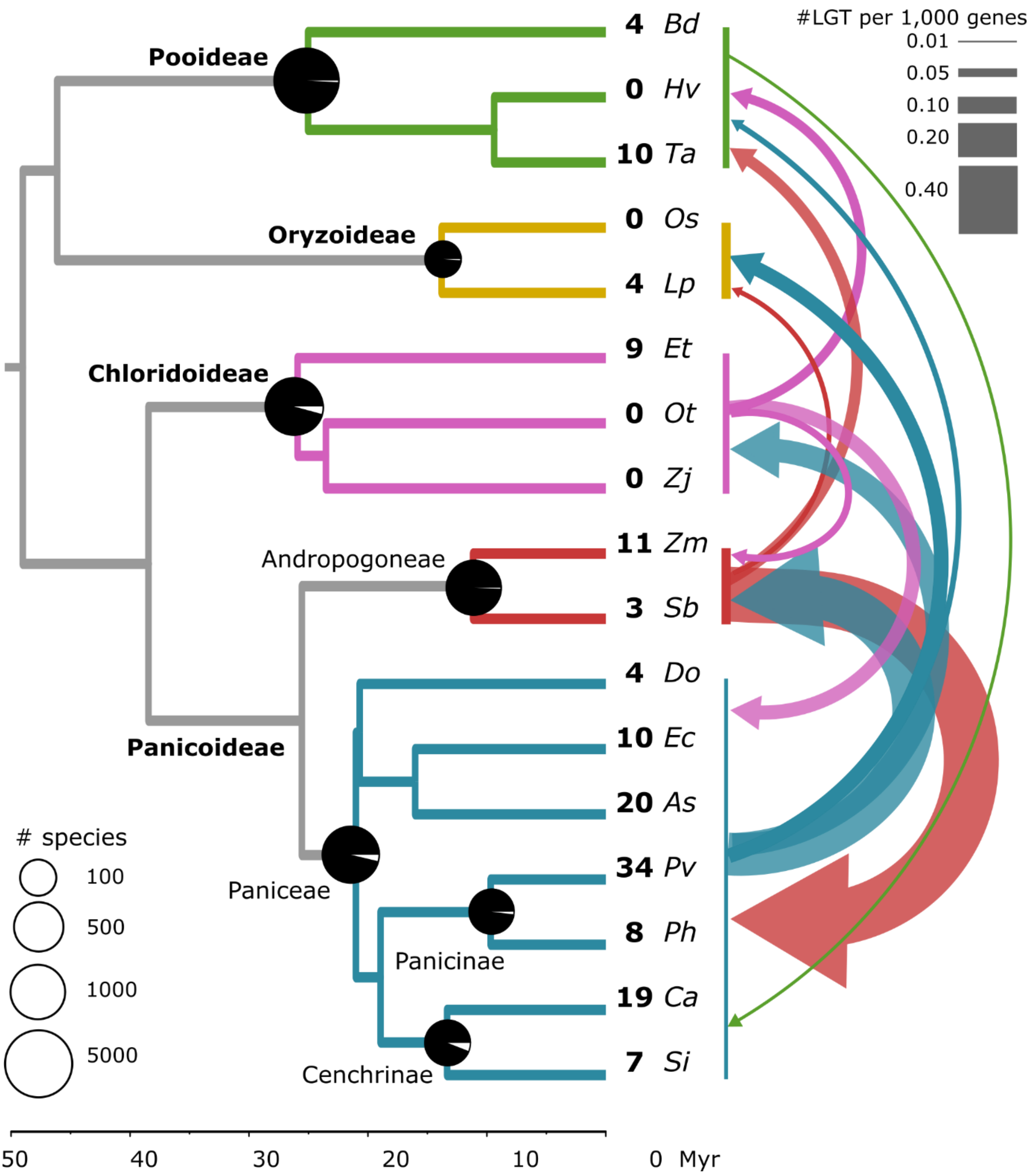
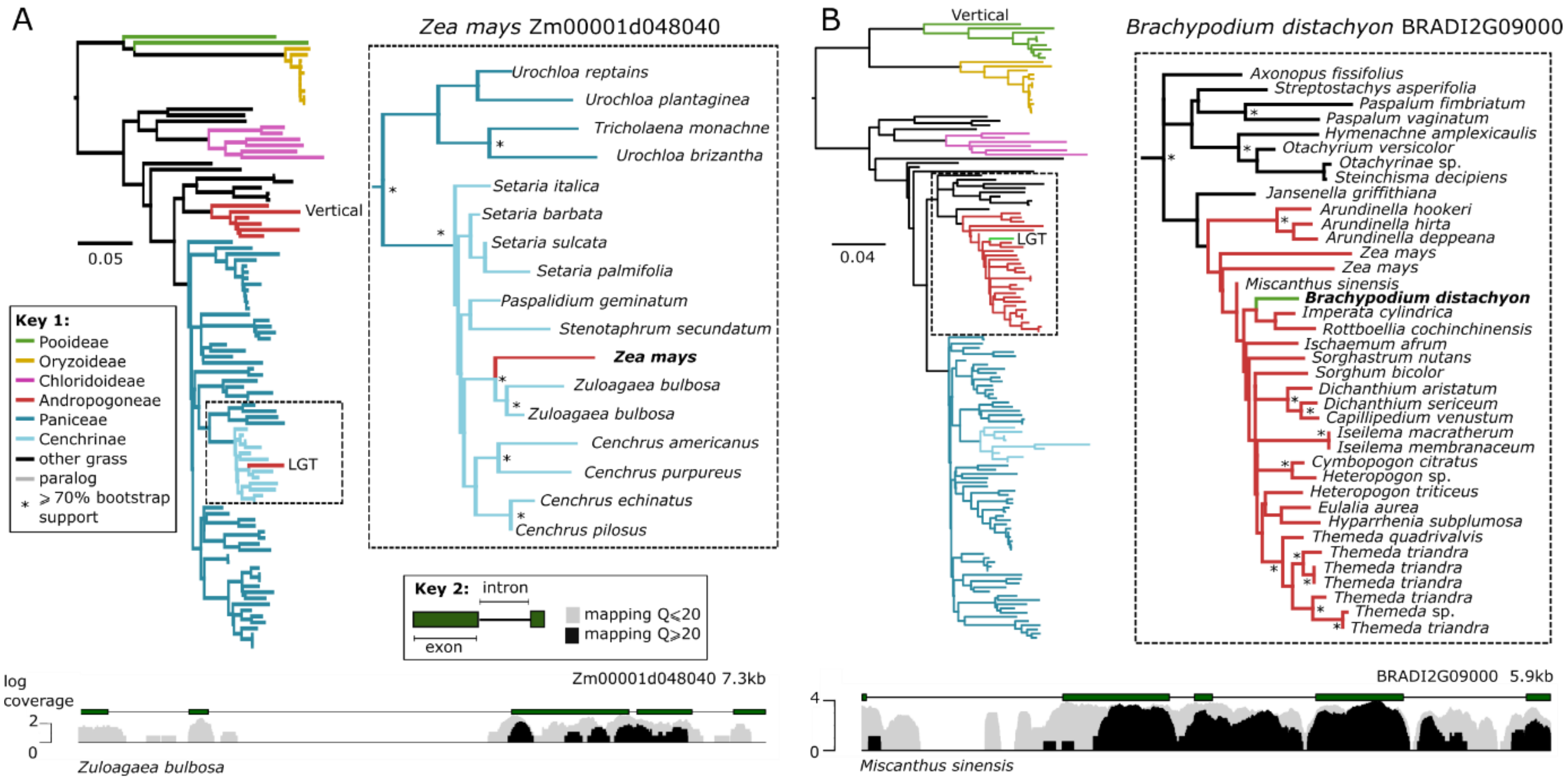


Figure 2.1: Distribution of lateral gene transfers (LGTs) among grasses. Time-calibrated phylogenetic tree of 17 grass species used in this study (phylogenetic tree from Christin *et al.*, 2014; scale in million years - Myr). The direction of LGT between grass clades is shown with arrows whose size is proportional to the number of LGTs received. The black portion of pie charts on key nodes of the phylogeny indicates the quartet support for the observed topology based on a multigene coalescence analysis (Dunning *et al.*, 2019). The size of each pie chart is proportional to the number of species within the clade (Soreng *et al.*, 2015). Numbers at the tips are the number of LGTs detected in each genome.

After these two successive filters, retained candidates were subjected to further validation. To verify the nesting of the candidate LGTs was not due to convergent adaptive amino acid substitutions, we generated phylogenetic trees based solely on 3rd codon positions, which are less subject to positive selection (Christin *et al*., 2012b). Phylogenetic trees were generated as above and were manually inspected to confirm the LGT scenario. To verify that the LGT scenario was statistically better than the species tree, we then conducted approximately unbiased (AU) topology tests that compared the maximum likelihood topology with a topology representing the null hypothesis (forcing monophyly of the donor and recipient clades; recipients for the within-Paniceae analysis were constrained at the genus level if they did not belong to the Cenchrinae or Panicinae). The null topology was inferred by first constraining the clades and inferring a tree with the GTR + G model in RaxML v.8.2.12 (Stamatakis, 2014), before using this topology as a constraint for a maximum likelihood phylogeny inferred with PhyML as described above. The AU tests were then performed in Consel v.1.20 (Shimodaira and Hasegawa, 2001) using the site-wise likelihood values generated by PhyML, and p-values were Bonferroni corrected to account for multiple testing. LGT candidates with non-significant results (p-value > 0.05) were discarded. In some cases, no native copy was present in any species from the group containing the focus species, preventing AU tests. These genes were retained, although the numbers were recorded separately (Table 2.2; n.b. statistics reported and values quoted in the text include these genes).

For candidates retained after these extra validation steps, new phylogenetic trees were inferred with a denser species sampling to refine the identification of the potential donor. Illumina short-read data sets (n = 71; 65 sp.; Supplementary Table 2.1) were added to the trees using the method described in Dunning *et al*., (2019). The dense trees were then manually inspected and any presenting strong discrepancies with the expected species relationships were discarded. All separate genes are counted in the final LGT tally for each species, so that duplicates (e.g. via polyploidization) arising after the transfer are counted separately (Supplementary Table 2.2).

In summary, to be considered as an LGT each gene (i) had to be nested within one of the other four groups of grasses (Figure 2.1); (ii) their nesting had to be well supported (≥ 70% bootstrap node support); (iii) potential parology problems had to be ruled out (i.e. discarding phylogenies with multiple apparent duplication events that can explain the phylogenetic incongruence); (iv) the nesting had to be supported by phylogenetic trees constructed solely from the 3rd codon positions, which are less subject to adaptive convergent evolution; and (v) where possible, the nesting had to be supported by approximately unbiased (AU) tests to confirm the LGT topology was a significantly better fit than a topology constrained to match the species tree (see Figure 2.2 for exemplar LGTs). Alignments (Supplementary Dataset 2.1) 85 taxa phylogenies (Supplementary Dataset 2.1, Supplementary Dataset 2.2), 3rd codon position phylogenies (Supplementary Dataset 2.3) and phylogenies with short-read data added (Supplementary Dataset 2.4) are included as supplementary datasets. All analyses were performed using publicly available data (Supplementary Table 2.1).


A
0.05
Vertical
LGT
Key 1:
Pooideae
Oryzoideae
Chloridoideae
Andropogoneae
Paniceae
Cenchrinae
other grass
paralog
* ⩾70% bootstrap support
Zea mays Zm00001d048040
Urochloa reptains
Urochloa plantaginea
Tricholaena monachne
Urochloa brizantha
Setaria italica
Setaria barbata
Setaria sulcata
Setaria palmifolia
Paspalidium geminatum
Stenotaphrum secundatum
Zea mays
Zuloagaea bulbosa
Zuloagaea bulbosa
Cenchrus americanus
Cenchrus purpureus
Cenchrus echinatus
Cenchrus pilosus
Key 2:
intron
exon
mapping Q⩽20
mapping Q⩾20
log coverage
2
0
Zm00001d048040 7.3kb
Zuloagaea bulbosa
B
Vertical
LGT
0.04
Brachypodium distachyon BRADI2G09000
Axonopus fissifolius
Streptostachys asperifolia
Paspalum fimbriatum
Paspalum vaginatum
Hymenachne amplexicaulis
Otachyrium versicolor
Otachyrinae sp.
Steinchisma decipiens
Jansenella griffithiana
Arundinella hookeri
Arundinella hirta
Arundinella deppeana
Zea mays
Zea mays
Miscanthus sinensis
Brachypodium distachyon
Imperata cylindrica
Rottboellia cochinchinensis
Ischaemum afrum
Sorghastrum nutans
Sorghum bicolor
Dichanthium aristatum
Dichanthium sericeum
Capillipedium venustum
Iseilema macratherum
Iseilema membranaceum
Cymbopogon citratus
Heteropogon sp.
Heteropogon triticeus
Eulalia aurea
Hyparrhenia subplumosa
Themeda quadrivalvis
Themeda triandra
Themeda triandra
Themeda triandra
Themeda triandra
Themeda sp.
Themeda triandra
BRADI2G09000 5.9kb
4
0
Miscanthus sinensis

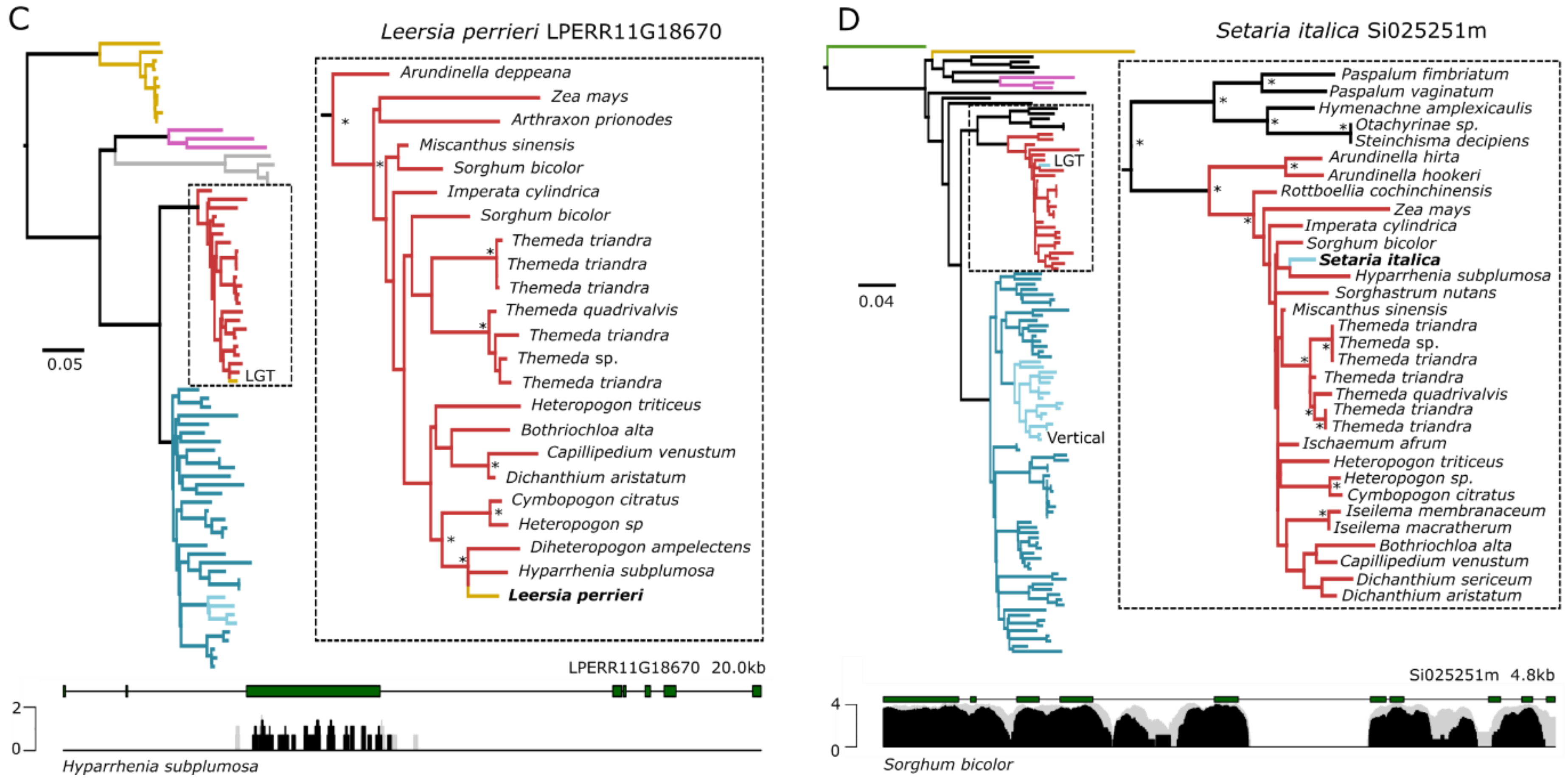


Figure 2.2: Four examples of grass-to-grass lateral gene transfer. Each panel (A-D) shows an exemplar grass-to-grass LGT, with full and expanded regions of maximum likelihood phylogenies shown. Asterisks denote nodes with bootstrap support values ≥70%, and branches are coloured per group. A coverage plot for each gene model is shown below, generated from short-read mapping data for a species closely related to the LGT donor.

2.3.2 Synteny analyses

Synteny analyses were performed with all genomes with reasonable contiguity (N50 ≥ 1Mb; n = 13; Supplementary Table 2.2) using SynFind (Tang *et al.*, 2015) with default parameters. For each LGT in these species, we determined whether genes from the other reference genomes identified as orthologs to the native copy in the phylogenetic trees were syntenic to the LGT or the native copy based on the highest syntelog score (Supplementary Table 2.3).

2.3.3 Analyses of replicate sequencing runs to check for potential contamination

Independently sequenced runs from the same accession or cultivar for each of the model species were screened for the presence of each LGT, as potential contaminations would not appear in multiple replicates derived from independent DNA samples. Paired-end Illumina whole-genome data were obtained from NCBI Sequence Read Archive and mapped to the reference genome using bowtie2 v.2.3.5.1 (Langmean & Salzberg, 2012) with default parameters. Mean coverage depths for the coding sequence of each gene in the genome were then calculated using bedtools v2.26.0 (Quinlan & Hall, 2010), with large bam files down-sampled with Picard Tools v.2.13.2-SNAPSHOT (Broad Institute, 2019).

2.3.4 Confirming LGT scenario with similarity of non-coding regions

Due to rapid divergence, non-coding sequences can only be accurately compared among close relatives. In the case of LGT, similarity of non-coding DNA is thus expected only when genome data are available for a close relative of the donor (see analyses of *A. semialata*; Dunning *et al.*, 2019; Olofsson *et al.*, 2019). We compared pairwise similarities of non-coding regions (intron and intergenic) of LGT regions versus the rest of the genome for a single multigene fragment from the *S. italica* genome. This fragment was selected as it has clear high-quality intergenic mapping (Q≥20) when using *S. bicolor* data as a proxy for the donor, suggesting that sequence data in this case are available for a close relative of the donor. For this analysis, paired-end Illumina whole-genome data belonging to the putative donor group, as well as close relatives of the recipient, were mapped to the reference genome as described above. We then used bedtools coverage to calculate the proportion of introns and intergenic regions with non-zero coverage with the different species, testing the hypothesis that coverage from the proxy donor is inflated around the putative LGTs. For introns, we restricted the analysis to those between 200bp and 2kb. For intergenic regions, we randomly generated windows using bedtools shuffle, excluding gene regions from the analysis.

2.3.5 Grass traits and statistical analyses

Plant traits were obtained from a variety of sources. Life history, distribution, growth form and the domestication status were retrieved from GrassBase (Clayton *et al.*, 2016). 1C Genome sizes were obtained from the Plant DNA C-values database (Pellicer & Leitch 2020), and climatic information from Watcharamongkol *et al.*, (2018). The climate data for *Oropetium thomaeum* were not included in Watcharamongkol *et al.*, (2018), and were therefore retrieved from GBIF [GBIF.org; 11th July (2019) GBIF Occurrence Download doi:10.15468/dl.wyhtoo] and WorldClim (Harris *et al.*, 2014; Fick & Hijmans, 2017) using the same methods. All statistical tests were performed in R v.3.0.2, with the expected frequencies for chi-square tests based on the number of genes tested within each species (Table 2.1). The Kruskal-Wallis tests were performed using absolute LGT numbers, which were divided into donor groups when testing whether some clades were more frequent donors than others. To determine if any trait or genome feature was associated with the number of LGTs, we performed phylogenetic generalized least squares (PGLS) to account for the relatedness between samples. The PGLS analysis was performed in R with the 'caper' package (Orme *et al.*, 2013) using a time-calibrated phylogenetic tree retrieved from Christin *et al.*, (2014), and various traits as explanatory variables (Table 2.1). Individual and iterative models were performed, removing the least significant variable until only significant variables remained (p-value <0.05).

2.4 Results

2.4.1 LGT occurs in all lineages and functional types of grass

Out of the 817,612 genes from the 17 grass genomes (Table 2.1) screened, 55.89% had sufficient homologous grass sequences (≥ 10 taxa) for reliable phylogenetic reconstruction (Table 2.2 & Supplementary Table 2.2), and were tested for LGT. A majority (99.73%) of the initial 37-taxa phylogenies did not support a scenario of LGT among the five grass groups, with successive filtering resulting in the identification of 135 LGT candidates across the 17 species (Table 2.2; full results Supplementary Table 2.2). Expectedly, a higher bootstrap threshold would decrease the number of retained candidates, but even a very conservative threshold of 95% support would identify 99 LGTs (Supplementary Figure 2.1). The number of LGTs received varied among species (p-value < 0.01; Chi-square test; mean = 8.4; SD=9.0; range=0 − 30; Supplementary Table 2.2), with the highest numbers observed in *Panicum virgatum* (n= 30), *Alloteropsis semialata* (n=20), and *Cenchrus americanus* (n=15). It should be noted that only a subset of the 59 previously reported LGTs in *Alloteropsis semialata* (Dunning *et al.*, 2019; Supplementary Table 2.4) are retrieved as the previous analysis examined additional groups of donors not considered here, and secondary candidates based solely on read-mapping patterns were not recorded in the present study. Despite the significant variation between species, the difference among the five phylogenetic groups was not significant (p-value = 0.16, Kruskal-Wallis test). Overall, our results show that LGT is widespread across the grass family and occurs in a majority of the species sampled here (Figure 2.1; Table 2.2). No LGTs were detected in four of the 17 species analysed, but some LGT might remain undetected due to our stringent phylogenetic filtering, and because we are only considering transfers among the predefined five grass clades.

Among the 17 species screened, LGT is observed in all functional groups (Figure 2.3). We detected LGT in wild species, but also in major crops. For instance, maize (*Zea mays*) has 11 LGTs received from Chloridoideae and Paniceae, while wheat (*Triticum aestivum*) has 10 LGTs received from Andropogoneae, Chloridoideae and Paniceae (Table 2). The LGTs may be beneficial for the crops, with transferred loci including some with functions related to abiotic stress tolerance and disease resistance (Supplementary Table 2.2). Across all plant properties, some seem associated with larger numbers of LGT (Figure 2.3). A phylogenetic generalized least squares (PGLS) analysis was conducted to test whether any of the traits had a significant relationship with the amount of LGT while accounting for phylogenetic effects. For this, we constructed a model to explain the absolute number of LGTs using nine traits as predictor variables (Table 2.1) and a time-calibrated phylogenetic tree retrieved from Christin *et al*., (2014). Initially, models were constructed for each predictor variable, with the amount of LGT shown to increase with the presence of rhizomes (p-value = 0.026, adjusted R2 = 0.243) and the number of genes tested (p-value = 0.038, adjusted R2 = 0.207). We subsequently performed a combined model with all explanatory variables to test for their joint effects. Iterative models were performed, removing the least significant variable until only significant variables remained (p-value <0.05). The PGLS analysis (combined adjusted R2 = 0.652) identified three characteristics that jointly explain the number of LGTs: the number of genes tested (p-value < 0.001), the presence of rhizomes (p-value = 0.002), and the ploidy level (p-value = 0.006). In the case of LGT happening prior to genome duplication, their number would be expected to double in tetraploids and triple in hexaploids because each homeologous chromosome will carry a copy of the LGT. Whilst we note that a majority of LGTs detected in the polyploids have been duplicated (n=43), there are still multiple singletons (n=35). These singletons were either acquired post-genome duplication, or they were possibly orphaned as a result of the complexity of polyploid genome assembly. Future studies should use larger sample sizes to definitely demonstrate the effects, but our analyses suggest that some categories of grasses are more likely to be involved in LGT.

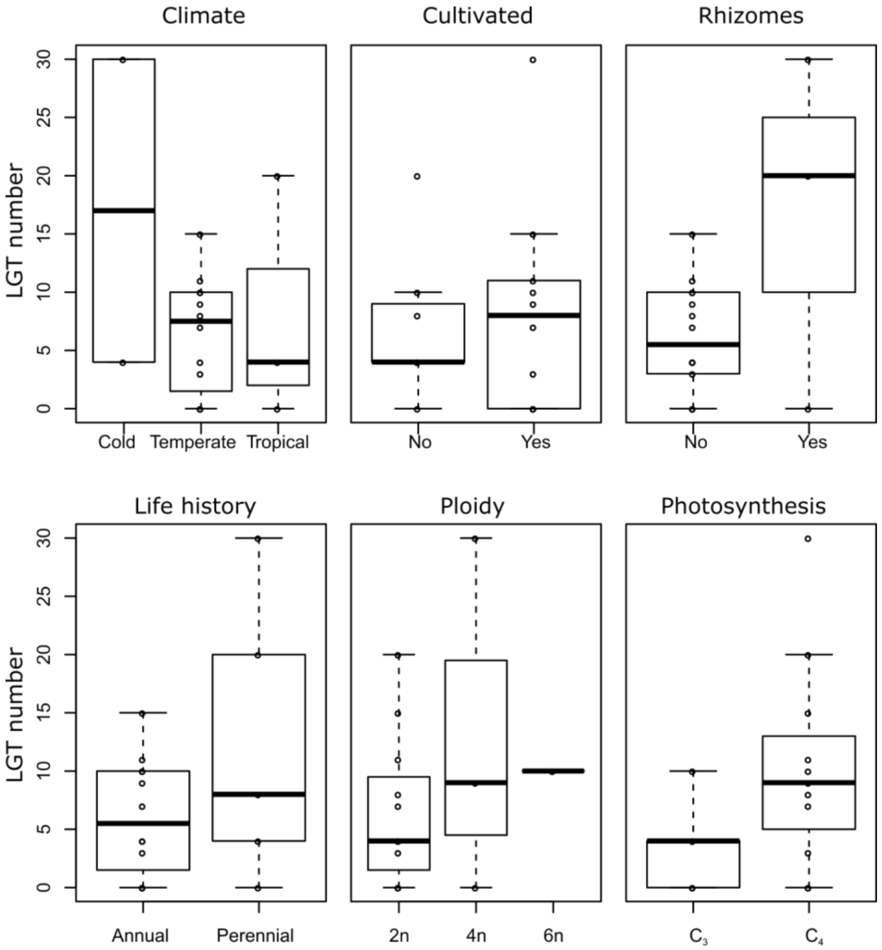


Figure 2.3: Number of lateral gene transfers (LGTs) received by different categories of grasses. For each group, the distribution of LGT numbers is shown with box plots connecting the median and the interquartile range, with whiskers showing 1.5 x the interquartile range. Individual data points are shown with dots.

2.4.2 LGTs are more commonly received from closely related species

Overall, some clades acted more frequently as donors (p-value < 0.01, Kruskal-Wallis test). Specifically, the Andropogoneae were the source of most transfers (Table 2.2). However, these were mainly received by members of Paniceae, which are the closest relatives of Andropogoneae in our dataset, and are also represented by the most genomes (Table 2.1). While these patterns suggest that LGT occurs more frequently among close relatives, directly comparing the rates is difficult because the clades vary in their number of species, number of genomes available and age. However, for a given clade of recipients, it is possible to compare the frequency of different groups of donors while controlling for their number of species. We therefore focused on the identity of donors of LGT to Paniceae, the group with the highest number of complete genomes from multiple genera.

Seven Paniceae genomes were used in this study, and this increased sample size further allows to detect intra-Paniceae LGT. We therefore reported the number of LGTs transferred from the Panicinae and Cenchrinae subgroups of Paniceae (each represented by two genomes; Figure 2.1) to other Paniceae, in addition to those received from other groups. In total, we identify 129 LGTs across the seven Paniceae genomes, 35 of which were transferred from the Cenchrinae and Panicinae subgroups (Table 2.3; full results Supplementary Table 2.5). When focusing on Paniceae recipients, some groups are more often LGT donors than others even after correcting for the number of species in each donor clade ($p < 0.01$, Kruskal-Wallis test). The number of LGTs given per species decreases with the phylogenetic distance to Paniceae, reaching lowest levels in the BOP clade (Pooideae and Oryzoideae; Figure 2.4).

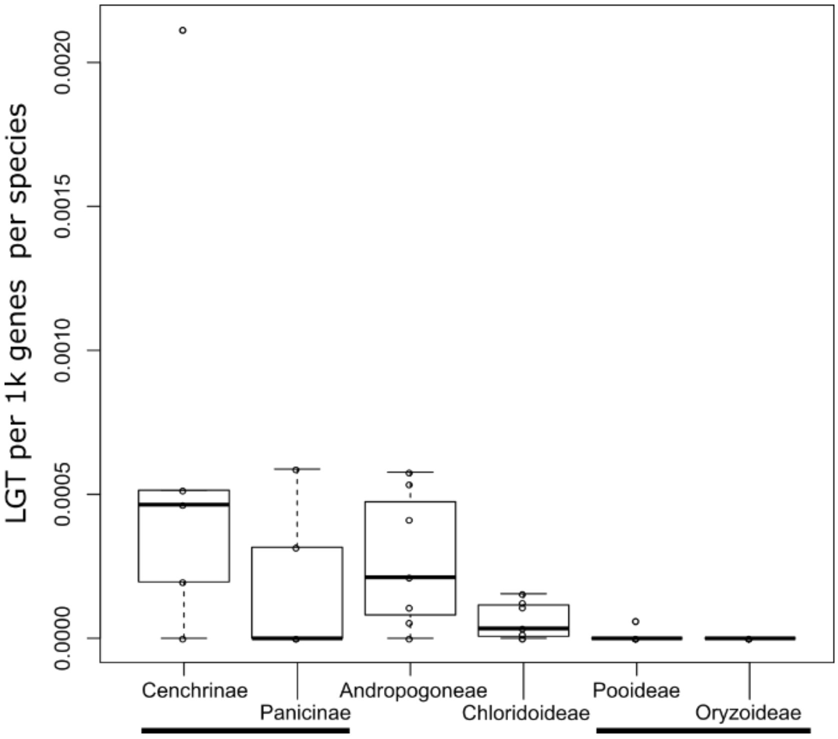


Figure 2.4 : Number of lateral gene transfers (LGT) received by Paniceae species from different groups. The number of LGTs in each Paniceae genome is corrected by the number of genes tested as well as the number of species in the group of donors. The phylogenetic distance increases from left to right, with equidistant clades joined by solid bars. Box plots show median, interquartile range and 1.5 x interquartile range, with individual data points shown with dots.

2.4.3 Ruling out alternative hypotheses

There are four main alternative hypotheses to LGT: [1] incomplete lineage sorting, [2] unrecognised parology, [3] hybridisation, [4] contamination, and [5] phylogenetic biases, such as convergent evolution. Below we present evidence reducing the likelihood of these alternative explanations.

[1] Incomplete lineage sorting: for a majority of the LGTs we detect (79.4%), the recipient genome also contains a native copy, which argues against incomplete lineage sorting as an alternative hypothesis. However, as pseudogenization of the native copy has been observed in cases where the LGT acts as a functional replacement (Dunning *et al*., 2019; Phansopa *et al*., 2020), their continued coexistence should not always be expected. The coexistence of native and laterally acquired orthologs permits us to compare patterns of synteny in multiple species to rule out unrecognised parology problems.

[2] Unrecognised parology: we used 13 species for this analysis, with at least two representatives from each of the five groups. For each LGT detected in these 13 species, we determined whether the genes from the other 12 species identified as orthologous in the phylogenetic tree were syntenic to the LGT or the native gene. In total, 76.2% of orthologs were syntenic with the native copy, 2.86% were syntenic with the LGT and 20.9% were syntenic to neither (Supplementary Table 2.3). The 2.86% of orthologs syntenic to the LGT correspond to three genes in *Echinochloa crus-galli* acquired from a Cenchrinae species and could result from technical (e.g. mis-assembly) or biological (e.g. homologous replacement) processes. Overall, the synteny analyses confirm that our phylogenetic trees identify true orthologs in most cases, and the phylogenetic patterns suggesting LGT cannot be explained by widespread unrecognised paralogy.

[3] Hybridisation: the patterns of synteny between the native and laterally acquired genes also argue against straightforward hybridisation through sexual reproduction and chromosomal recombination during the transfers, as already argued previously (Dunning *et al*., 2019). With the exception of three genes in Echinochloa crus-galli, the LGTs appear to be inserted into the genome in random locations, often on different chromosomes as the native orthologs.

[4] Contamination: we rule out contamination as the source of the foreign DNA in the genomes by confirming the presence of the laterally acquired DNA in multiple independent sequencing runs (Supplementary Table 2.6 and Supplementary Figure 2.2). For six of the reference genomes, 'gold-standard' datasets exists, i.e. whole-genome resequencing data sets for the same cultivar as the reference genome, but that were produced independently from the initial assembly project (Supplementary Table 2.6). A further four genomes had multiple libraries from the original assembly project, and these were derived from independent DNA samples (Supplementary Table 2.6). For the remaining three species, the available sequencing data cannot be used to rule out contamination as only the whole-genome data

used to generate the reference assembly exists, and where there are multiple sequencing libraries/runs it is unclear whether they are derived from independent DNA samples (Supplementary Table 2.6). For each dataset, we compared the genome-wide mean per-base coverage for each gene to that of the identified LGT (Supplementary Table 2.6 and Supplementary Figure 2.2), with an expectation that a gene corresponding to sample contamination would have zero (or near zero) coverage in all but one independently produced sequencing runs used to assemble the reference genome, and in none of the sequencing runs produced independently of the reference genome. All LGTs had sequencing data in all independent datasets apart from one gene in Z. mays. For this species, we used seven datasets from the same cultivar that were produced independently in seven different labs. Only five out of these seven datasets supported the presence of the LGT Zm00001d039537, with the most parsimonious explanation being LGT variation between individuals, as previously documented in *Alloteropsis semialata* (Dunning *et al*., 2019). A majority of LGTs had coverage depths greater than the 5th (97.0% of LGTs) and 2.5th (99.0% of LGTs) percentile of coverage depth for all genes in the genome (Supplementary Table 2.6 and Supplementary Figure 2.2). Overall, these results confirm that, at least for species with independent replicates, contamination in the original reference genomes is not responsible for the presence of the LGTs in the sequence datasets.

[5] Phylogenetic bias: convergent evolution or other systematic biases in the data could lead to gene/species tree discordance (Chang & Campbell, 2000). In addition to confirming the patterns with phylogenetic trees built on third positions of codons, we assessed the similarity between the recipient and donor species in non-coding DNA. The mapping of short-read data to four genomes confirmed in some cases a high similarity between the putative donor and recipient on intron sequences of LGTs in addition to exons (Figure 2.2). It was however not possible to delimit with high precision the laterally acquired fragments detected here (as done for *A. semialata* in Dunning *et al*., 2019 and Olofsson *et al*., 2019), either because the transfers are too ancient or because we lack whole genome data for very close relatives of the donors. However, we did detect a multigene fragment in *Setaria italica* that also appeared to have laterally acquired intergenic DNA when using *Sorghum bicolor* mapping data as a proxy for the unknown Andropogoneae donor (Supplementary Figure 2.3). For this fragment, we quantified the mapping rates between intron and intergenic LGT regions to the rest of the genome. Out of the 20,972 genes from *S. italica* with at least one intron between 200 and 2,000 bp, only 164 had a higher proportion of bases covered by *S. bicolor* reads than the three LGTs in the *S. italica* fragment. Of these 164 genes, only 67 were covered by more reads of the species from the donor group (*S. bicolor*) than of the close relative of *S. italica* that is *Cenchrus americanus*. The multi-gene fragment also includes 4.5kb of laterally acquired intergenic DNA with 91.3% non-zero coverage with the *S. bicolor* data (Supplementary Figure 2.3). We compared this to 10,000 other randomly sampled 4.5kb intergenic regions across the genome, all of which had a lower non-zero coverage than that of the LGT region

(mean = 2.6%; SD = 7.1%). The observation of some intergenic regions with high similarity (Supplementary Figure 2.3), together with intronic similarities (Supplementary Figure 2.3), further rules out convergent evolution or other phylogenetic biases (e.g. long branch attraction) as being responsible for all detected cases of gene/species tree discordance.

2.5 Discussion

Lateral gene transfer is a potent evolutionary force capable of having a profound impact on the evolutionary trajectory of a species and its descendants (Li *et al.*, 2014; Cheng *et al.*, 2019; Phansopa *et al.*, 2020; Chen *et al.*, 2021). Here, we use grasses as a model to investigate the factors that dictate the prevalence of LGT among plants. Using a combination of stringent phylogenetic and genomic analyses, we have identified a grand total of 170 genes (approximately 3.72 LGTs per 10,000 genes; 135 between the five large groups of grasses and 35 among groups of Paniceae) that have been laterally transferred to 13 of the 17 complete grass genomes that were screened (Table 2.1 & Table 2.3). Our approach was developed to drastically reduce the amount of false positives, and is purposely very conservative. This enables us to minimise the effects of other evolutionary processes such as hybridisation and incomplete lineage sorting. As a result, the number of LGTs identified is likely only a subset of those existing in the complete grass genomes. In addition, the phylogenetic filtering prevents us from detecting LGT from clades of grasses for which no genome is available. With the current sampling, at least 30% of the grass diversity is never considered as potential LGT donors (Soreng *et al.*, 2015). The number of detected LGTs therefore depends on the sampling of genomes, and future studies with more species representing additional potential donors will likely lead to more LGT discoveries. Our efforts already indicate that the phenomenon is prevalent in the family.

Our phylogenetic pipeline prevents us from detecting LGT happening among members of the same group of grasses, such as the numerous exchanges among lineages of Paniceae previously detected (Dunning *et al.*, 2019). This is perfectly exemplified by the case of *A. semialata*, in which 26 LGTs were previously detected based on phylogenetic analyses (referred to as ‘primary LGT’, with 33 ‘secondary candidates’ detected based on similarity in flanking regions in Dunning *et al.*, 2019; Supplementary Table 2.4). Here, only 20 were identified when considering solely LGT among the five higher groups (Table 2.2), while a further 14 were detected when considering subgroups of Paniceae as potential donors (Table 2.3). Seven more LGT were previously detected in *A. semialata* from a group of Paniceae (Melinidinae) that was not considered as putative donors here because of the absence of reference genomes (Supplementary Table 2.4). These differences highlight the influence of the availability of genomes for putative donors on our ability to detect LGT. In addition, five of the 34 LGTs detected here in the genome of *A. semialata* were not identified in the previous analysis of the same genome, showing that LGT detection depends on multiple factors. On the one hand, the removal of a first filter based on similarity analyses in the present study allowed identifying additional LGTs (Supplementary Table 2.4). On the other hand, the increased number of genomes in the present study influences the correction of p-values for multiple testing, leading in some cases to non-significant topology tests (Supplementary Table 2.4). In addition, the detection of LGT candidates based on secondary screening of flanking regions in the previous study (33 ‘secondary LGT candidates’ in Dunning *et al.*, 2019; Supplementary Table 2.4) demonstrates that some LGTs cannot be identified based solely on phylogenetic analyses, because they are too short or not present in enough species to infer robust phylogenetic trees. Finally, our approach precludes the detection of older LGTs that are shared by multiple species among the 17 reference genomes, as reported in other cases (e.g. Li *et al.*, 2014). We conclude that the LGTs we report here concern only a small fraction of those existing in grass genomes. Despite these limitations, we show that LGT is common in grasses, certain groups exchange more genes than others, the frequency of LGT appears to increase in rhizomatous species, and there may be a role of phylogenetic distance underpinning the LGT dynamics. Analyses based on more genomes will in the future refine our conclusions, and potentially provide more statistical power to precisely quantify the effect of different factors on the rate of LGT.

2.5.1 LGT occurs in all functional groups, and is especially prevalent in rhizomatous species

LGT is common in grasses and is observed in each of the five groups investigated here (Figure 2.1). We detected LGT in domesticated and wild species alike (Figure 2.3), although it is currently unknown whether the LGTs occurred before or after domestication and whether these genes are associated with agronomic traits. The genetic exchanges are not restricted to any functional category of grasses (Figure 2.3), and the ubiquity of the phenomenon provides some support for a breakdown in reproductive behaviour and illegitimate pollination as the mechanism responsible for the transfers as wind pollination is universal in this group. Further work is required to determine how traits associated with the wind-pollinated syndrome (e.g. self-compatibility, plant height and pollen longevity) could affect LGT among grasses. There is also a statistical increase of the number of LGTs in rhizomatous species and two of the three species with the highest numbers of LGTs (*Alloteropsis semialata* and *Panicum virgatum*) are perennials that can propagate vegetatively via rhizomes (Table 2.1 & Table 2.3). These patterns suggest that root-to-rhizome contact (i.e. inosculation) provides an increased opportunity for retaining gene transfers, as the integration of foreign DNA in rhizome tissue means that any subsequent plant material regrown from these cells, including reproductive tissue, will contain the LGT. This hypothesis is compatible with previous reports of genetic exchanges following grafts (Stegemann & Bock, 2009; Hertle *et al.*, 2021). In this instance, LGT is similar to somatic mutations occurring in clonal species, as documented in the seagrass *Zostera marina* where they can ultimately enter the sexual cycle (Yu *et al.*, 2020). The genetic bottleneck and selection characterising rhizomes would further increase the chance of LGT retention, especially if these provide a selective advantage (Yu *et al.*, 2020).

However, we did not detect LGT in the third rhizomatous species we sampled (*Zoysia japonica*; Table 2.1). Increased species sampling, particularly for rhizomatous species represented by only three genomes in this study, is now needed to confirm our conclusions and precisely quantify the impact of growth form on the amount of gene transfers and how it interacts with other factors.

2.5.2 It is easier to acquire genes from close relatives

Within grasses, there is an effect of the phylogenetic distance on the number of transfers observed, as shown by the Paniceae receiving more LGTs from closer relatives (Figure 2.4). This pattern mirrors that observed in prokaryotes (Popa & Dagan, 2011; Skippington & Ragan, 2012; Soucy *et al.*, 2015) and insects (Peccoud *et al.*, 2017), where the frequency of transfers is higher between closely related species. In prokaryotes, this effect is thought to result from more similar DNA sequence promoting homologous replacement of the native copy (Skippington & Ragan, 2012). This is unlikely to play a role in grasses as the LGTs are predominately inserted in non-syntenic positions in the genome where they coexist with the native copy (Supplementary Table 2.3). However, stretches of DNA similar between the donor and recipient (e.g. transposable elements) may still be involved in the incorporation of the LGT onto the chromosomes, a hypothesis that can be tested when genome assemblies for donor species become available. Alternatively, the effect of the phylogenetic distance might stem from the regulation of the LGT post acquisition, with genes transferred from closely related species more likely to share regulatory mechanisms. In such a scenario, the phylogenetic effect would reflect the utility of the LGT for the recipient species and therefore selection after the transfer rather than the rate of transfer. Overall, our analyses indicate that it is easier to either obtain LGTs from close relatives or to use it after the transfers, thereby increasing the chance of selectively retaining it.

2.5.3 A potential role of overlapping distributions.

We observe some transfers between Pooideae and Paniceae, two groups that diverged >50 Ma, representing one of the earliest splits within this family (GWPGII, 2012). This indicates that LGT is possible across the whole grass family. In our dataset, the only recipient of these transfers is *Dichanthelium oligosanthes* (Table 2.2), a frost-tolerant grass from North America that inhabits colder areas than other members of the Paniceae (Studer at al., 2016). In cold regions, *D. oligosanthes* can co-occur with members of the Pooideae, and this biogeographic pattern likely facilitated exchanges between the two groups of grasses. However, given the difficulties of identifying the donor to the species level (or even genus) with the current data, we cannot be sure that the specific donor and *D. oligosanthes* co-occur. As more whole-genome datasets become available for the diverse Pooideae, co-occurrence between the donor and recipient species can be directly tested.

Biogeography might also be responsible for differences in the identity of the LGT donors between the two closely related *Panicum* species. Indeed, a majority (75%) of LGTs in *Panicum hallii* were received from Chloridoideae, while a majority (81%) of those in *Panicum virgatum* were received from Andropogoneae (Table 2.3). This pattern mirrors the dominant grassland type (Chloridoideae vs. Andropogoneae) for a majority of the range of each of the two species, and the area from which the individual for the genome assembly was sampled (Lovell *et al*., 2018; Lehmann *et al*., 2019).

Quantifying the effects of biogeography as opposed to other factors requires identifying the donor to the species level and a detailed description of the spatial distribution of each grass species, including their abundances. Indeed, the likelihood of encounters will increase with the number of individuals of the donor species and not just its presence. In addition, the scale of relevant interactions would depend on the transfer mechanisms, with pollination- or vector-mediated transfers potentially able to move genes across plants from a given region, while direct transfers between plants (e.g. via inosculation) would only happen among directly adjacent species. Detailed ecological datasets coupled with genomic data for a large numbers of species are therefore needed to precisely assess the effect of biogeography on LGT dynamics in grasses.

2.5.4 Conclusion

Using stringent phylogenomic filtering, we show here that lateral gene transfer (LGT) is a widespread process in grasses, where it occurs in wild species as well as in widely cultivated crops (e.g. maize and wheat). LGT does not appear restricted to particular functional types, although it seems to increase in rhizomatous species, where vegetative growth offers extra opportunities for gene transfers into the germline. In addition, we show that the amount of successful transfers decreases with phylogenetic distance. This effect of the phylogenetic distance might result from increased genomic compatibility among more related groups. Thanks to the rapid accumulation of genome data for various groups of grasses, future studies of LGT will be able to sample densely the diversity of grasses and therefore refine our conclusions. However, with the current data we show that LGT occurs in a variety of grasses, highlighting the potential impact of the frequent movement of functional genes between species on the evolution of this critical group of plants.

2.5.5 Acknowledgements

This work was funded by a European Research Council grant (ERC-2014-STG-638333) and a Royal Society Research grant (RGF\EA\180247). P.-A.C. is supported by a Royal Society University Research Fellowship (URF\R\180022). L.T.D is supported by a Natural Environment Research Council Independent Research Fellowship (NE/T011025/1).

2.6 Tables

Table 2.1: Species used in this study and associated traits.

Group	Species	Ploid.	1C	#Genes tested	Cult.	LH	Clim.	Phot.	Cont.	Rhiz.
Pooideae	*Brachypodium distachyon*[1]	2n	0.31	17204	N	A	Temp	C_3	6	N
Pooideae	*Hordeum vulgare*[2]	2n	5.39	16192	Y	A	Temp	C_3	6	N
Pooideae	*Triticum aestivum*[3]	6n	16.95	56619	Y	A	Temp	C_3	6	N
Oryzoideae	*Oryza sativa*[4]	2n	0.49	19259	Y	A	Trop	C_3	6	N
Oryzoideae	*Leersia perrieri*[5]	2n	0.32	15777	N	A	Trop	C_3	1	N
Chloridoideae	*Eragrostis tef*[6]	4n	0.69	30605	Y	A	Temp	C_4	5	N
Chloridoideae	*Oropetium thomaeum*[7]	2n	0.29	15168	N	A	Temp	C_4	2	N
Chloridoideae	*Zoysia japonica*[8]	4n	0.42	20416	Y	P	Temp	C_4	1	Y
Andropogoneae	*Sorghum bicolor*[9]	2n	0.69	21962	Y	A	Temp	C_4	6	N
Andropogoneae	*Zea mays*[10]	2n	2.65	25866	Y	A	Temp	C_4	6	N
Paniceae	*Alloteropsis semialata*[11]	2n	1.10	23071	N	P	Trop	C_4	3	Y
Paniceae	*Cenchrus americanus*[12]	2n	2.65	20159	Y	A	Temp	C_4	4	N
Paniceae	*Dichanthelium oligosanthes*[13]	2n	0.96	17761	N	P	Cold	C_3	1	N
Paniceae	*Echinochloa crus-galli*[14]	6n	1.37	54181	N	A	Temp	C_4	6	N
Paniceae	*Panicum hallii*[15]	2n	0.55	30255	N	P	Temp	C_4	1	N
Paniceae	*Panicum virgatum*[16]	4n	1.89	45043	Y	P	Cold	C_4	3	Y
Paniceae	*Setaria italica*[17]	2n	0.49	27465	Y	A	Temp	C_4	6	N

Ploid. = Ploidy; 1C = 1C genome size in Gb; Cult. = cultivated (Y = yes; N = no); LH = life history (A = annual; P = perennial); Clim. = climate (Temp = temperate; Trop = tropical); Phot. = photosynthetic type; Cont. = number of continents; Rhiz. = rhizomatous (Y = yes; N = no). [1]International Brachypodium Initiative, 2010; [2]International Barley Genome Sequencing Consortium, 2012; [3]International Wheat Genome Sequencing Consortium, 2014; [4]Goff *et al.*, 2002; [5]Stein *et al.*, 2018; [6]Cannarozzi *et al.*, 2014; [7]VanBuren *et al.*, 2015; [8]Tanaka *et al.*, 2016; [9]Patterson *et al.*, 2009; [10]Schnable *et al.*, 2009; [11]Dunning *et al.*, 2019; [12]Varshney *et al.*, 2017; [13]Studer *et al.*, 2016; [14]Guo *et al.*, 2017; [15]Lovell *et al.*, 2018; [16]*Panicum virgatum* v4.1, DOE-JGI, http://phytozome.jgi.doe.gov/; [17]Bennetzen *et al.*, 2012

Table 2.2: Number of lateral gene transfers (LGT) detected between the five groups.

			Donor clade				
Clade	**Species**	**# LGT**	**Pooid.**	**Ory.**	**Chlor.**	**Andro.**	**Pan.**
Pooideae	*Brachypodium distachyon*	4	-	0	0	4	0
Pooideae	*Hordeum vulgare*	0	-	0	0	0	0
Pooideae	*Triticum aestivum*	8(10)	-	0	5	0(2)	3
Oryzeae	*Oryza sativa*	0	0	-	0	0	0
Oryzeae	*Leersia perrieri*	1(4)	0	-	0	1	0(3)
Chloridoideae	*Eragrostis tef*	1(9)	0	0	-	0	1(9)
Chloridoideae	*Oropetium thomaeum*	0	0	0	-	0	0
Chloridoideae	*Zoysia japonica*	0	0	0	-	0	0
Andropogoneae	*Sorghum bicolor*	2(3)	0	0	0	-	2(3)
Andropogoneae	*Zea mays*	11	0	0	2	-	9
Paniceae	*Alloteropsis semialata*	20	0	0	4	16	-
Paniceae	*Cenchrus americanus*	15	0	0	5	10	-
Paniceae	*Dichanthelium oligosanthes*	4	4	0	0	0	-
Paniceae	*Echinochloa crus-galli*	10	0	0	3	7	-
Paniceae	*Panicum hallii*	8	0	0	6	2	-
Paniceae	*Panicum virgatum*	30	0	0	1	29	-
Paniceae	*Setaria italica*	7	0	0	0	7	-

The numbers in parentheses include genes for which approximate unbiased (AU) topology tests could not be performed as no native copy from the same clade was present to constrain the tree topology. Pooid. = Pooideae; Ory. = Oryzoideae; Chlor. = Chloridoideae; Andro. = Andropogoneae; Pan. = Paniceae.

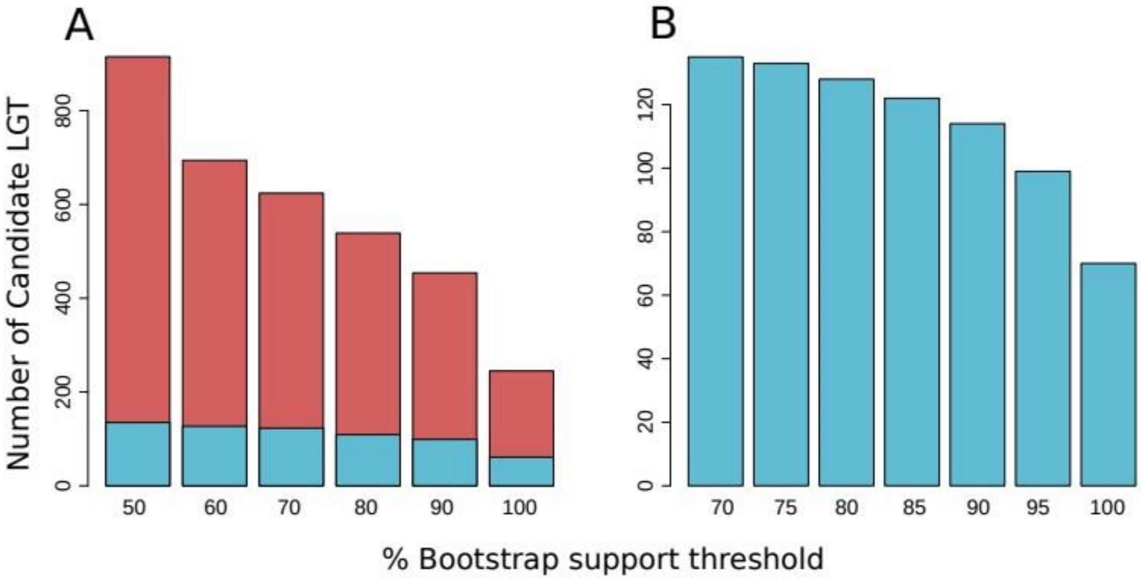
Table 2.3: Number of lateral gene transfers (LGT) detected in Paniceae.

Subgroup	**Species**	**# LGT**	**Pooid. (3,698 sp.)**	**Ory. (115 sp.)**	**Chlor. (1,602 sp.)**	**Andro. (1,202 sp.)**	**Cench. (287 sp.)**	**Pani. (157 sp.)**
Cenchrinae	*Cenchrus americanus*	16	0	0	5	10	-	1
Cenchrinae	*Setaria italica*	7	0	0	0	7	-	0
Panicinae	*Panicum hallii*	8	0	0	6	2	0	-
Panicinae	*Panicum virgatum*	36	0	0	1	29	6	-
Other	*Alloteropsis semialata*	33(34)	0	0	4	16	13(14)	0
Other	*Dichanthelium oligosanthes*	5	4	0	0	0	1	0
Other	*Echinochloa crus-galli*	23	0	0	3	7	8	5

The number of species in each clade is indicated in parentheses, with values from Soreng *et al.*, (2015); Pooid. = Pooideae; Ory. = Oryzoideae; Chlor. = Chloridoideae; Andro. = Andropogoneae; Cench. = Cenchrinae; Pani. = Panicinae.

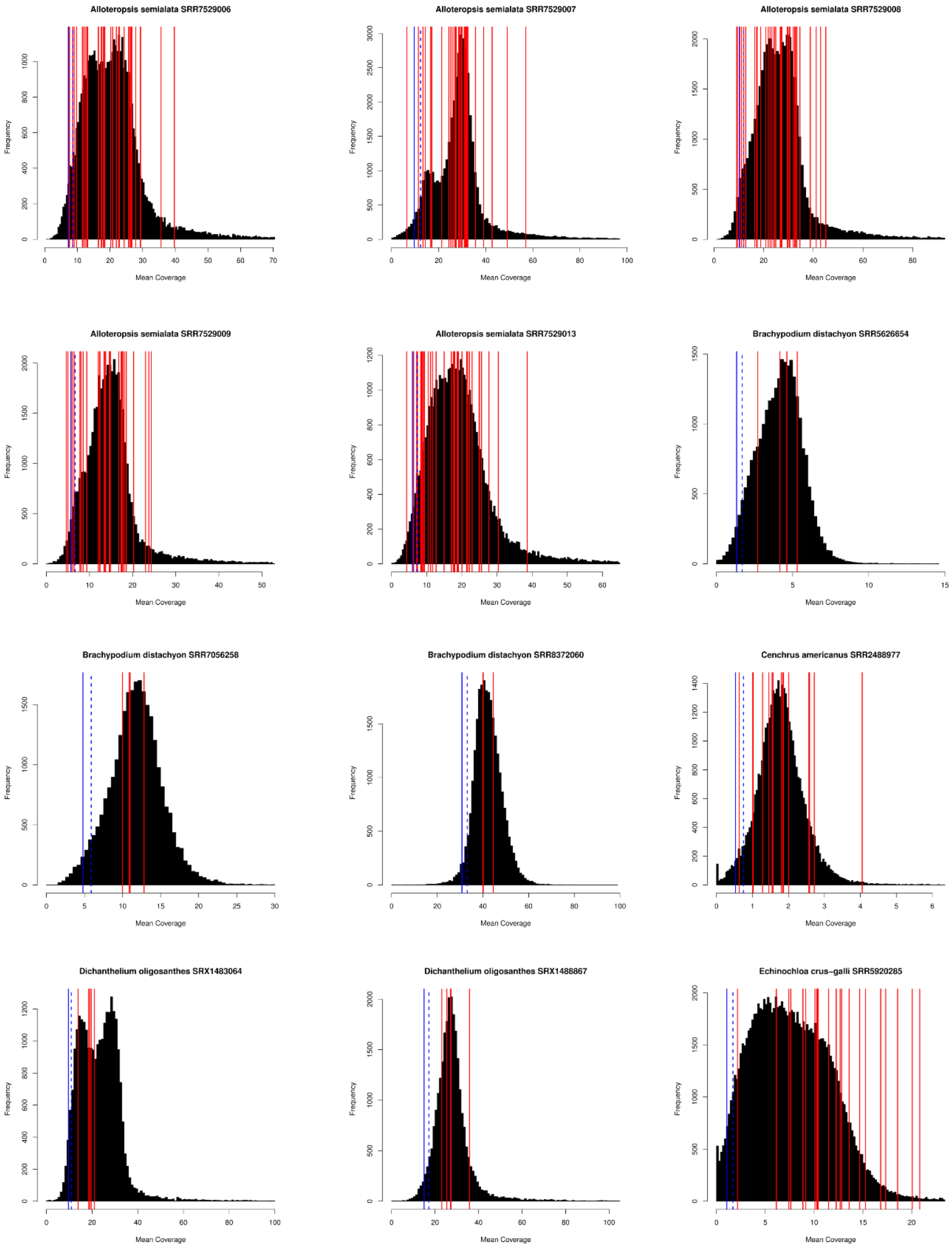
2.7 Supplementary Information

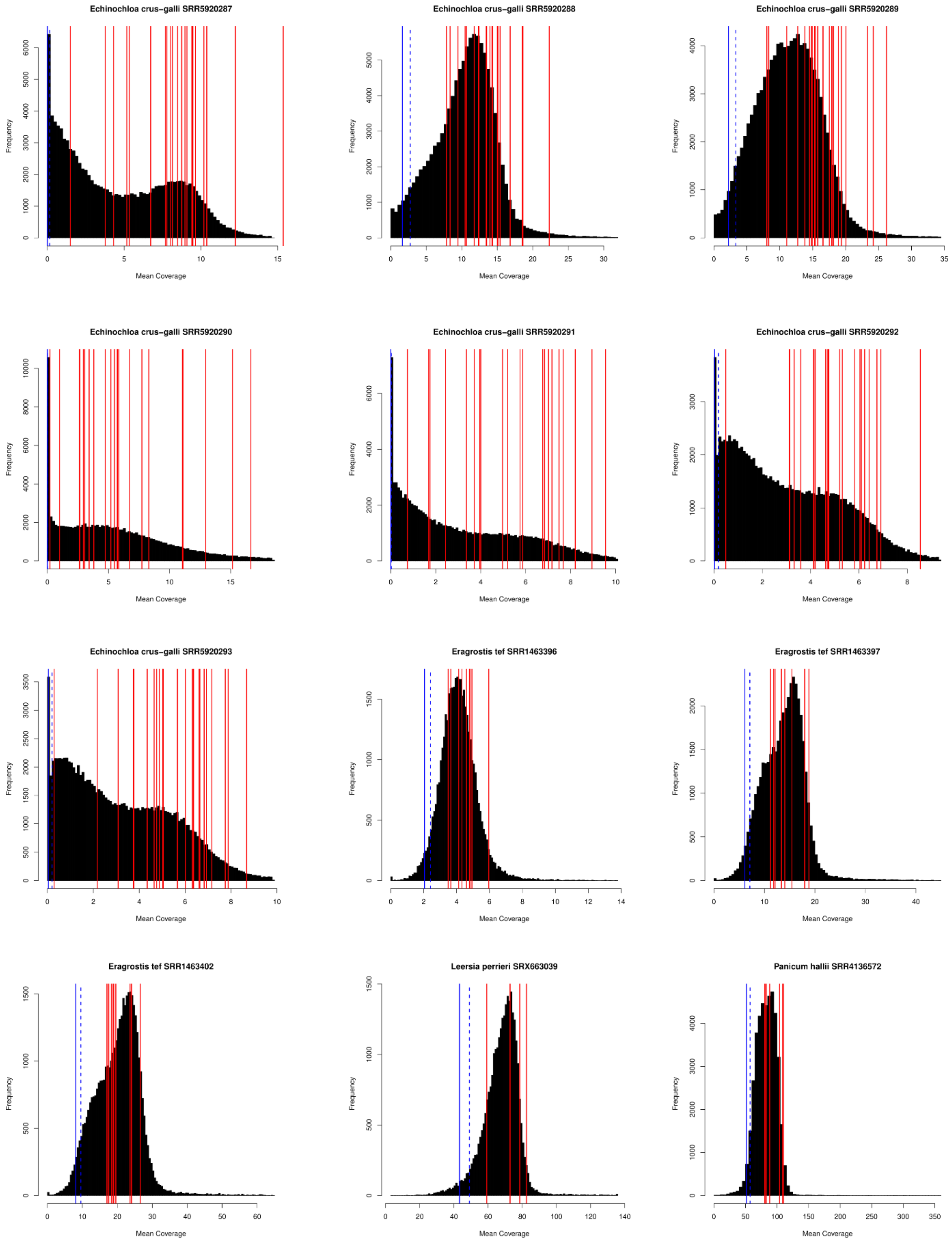
2.7.1 Supplementary Figures



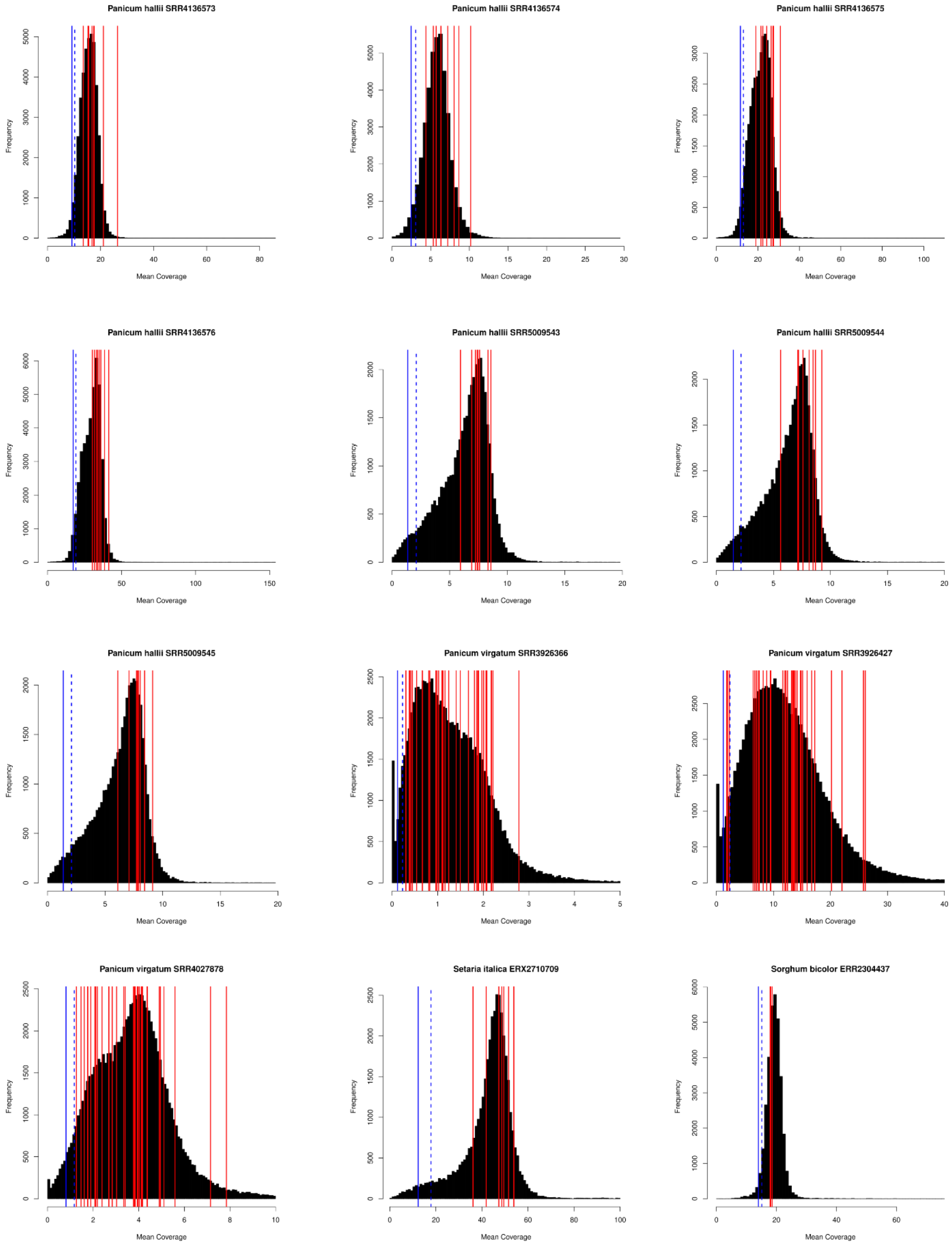
Supplementary Figure 2.1: Impact of varying filtering parameters on LGT detection. The number of LGT detected in the main 17 species analysis is shown for various bootstrap support thresholds. Panel A shows the first filtering step on the 37 taxa trees and panel B shows the second filtering step on the 85 taxa trees. The number of candidates passing the first filter are indicated with red bars, and the final number of LGT (with a threshold during the second filter of 70% in panel A) is shown with blue bars.

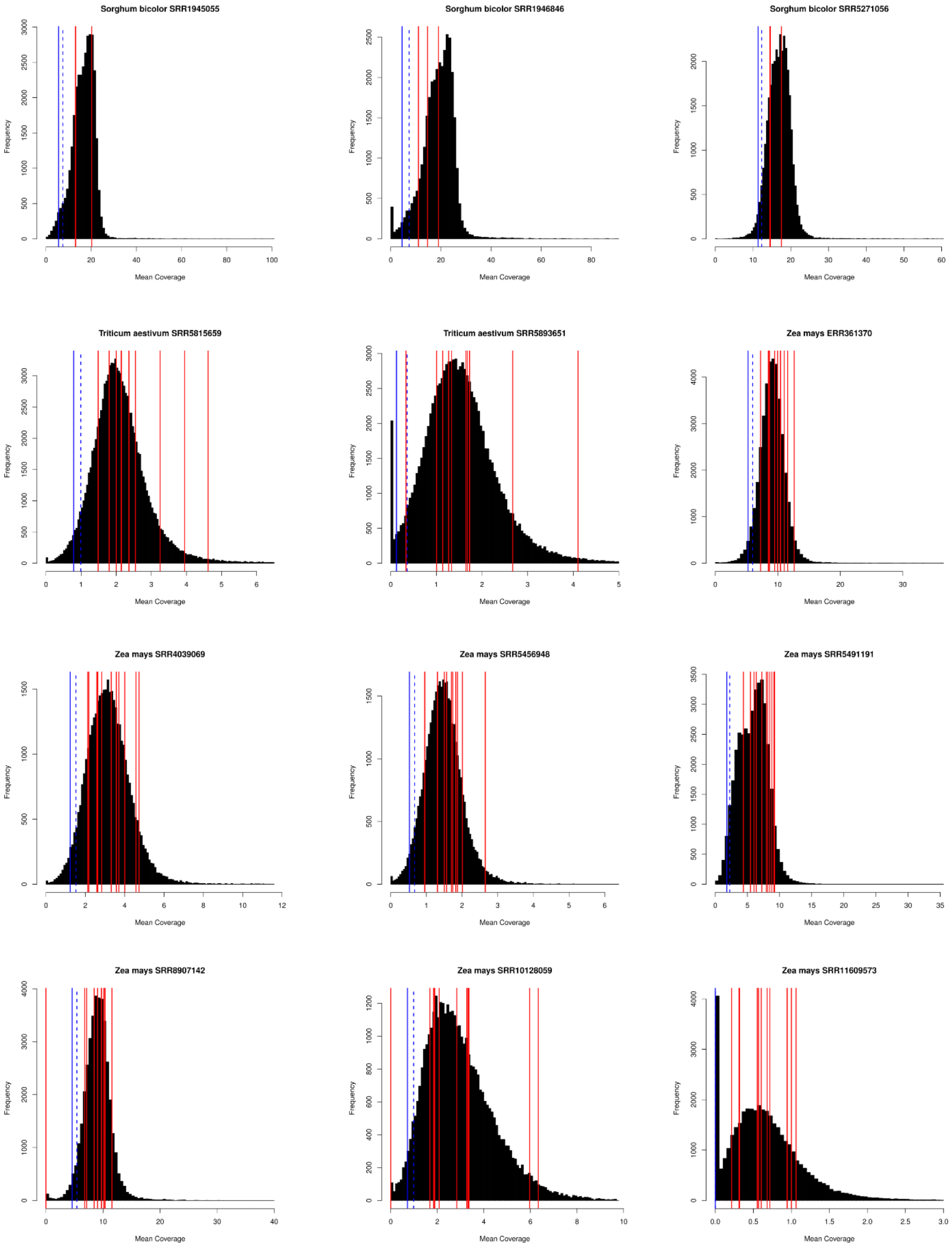
Supplementary Figure 2.2: Coverage plots comparing independent sequencing runs. To evaluate the likelihood that the detected patterns are due to contamination, different NCBI Sequence Read Archive datasets were analysed. In each case, the species and accession number are indicated, and the black histogram shows the distribution of mean per-base coverage depths across all genes from the genome. Solid blue lines show the 2.5th percentile and dashed blue lines show the 5th percentile, while red lines indicate the mean per-base coverage depth for each of the LGTs detected in the reference genome.



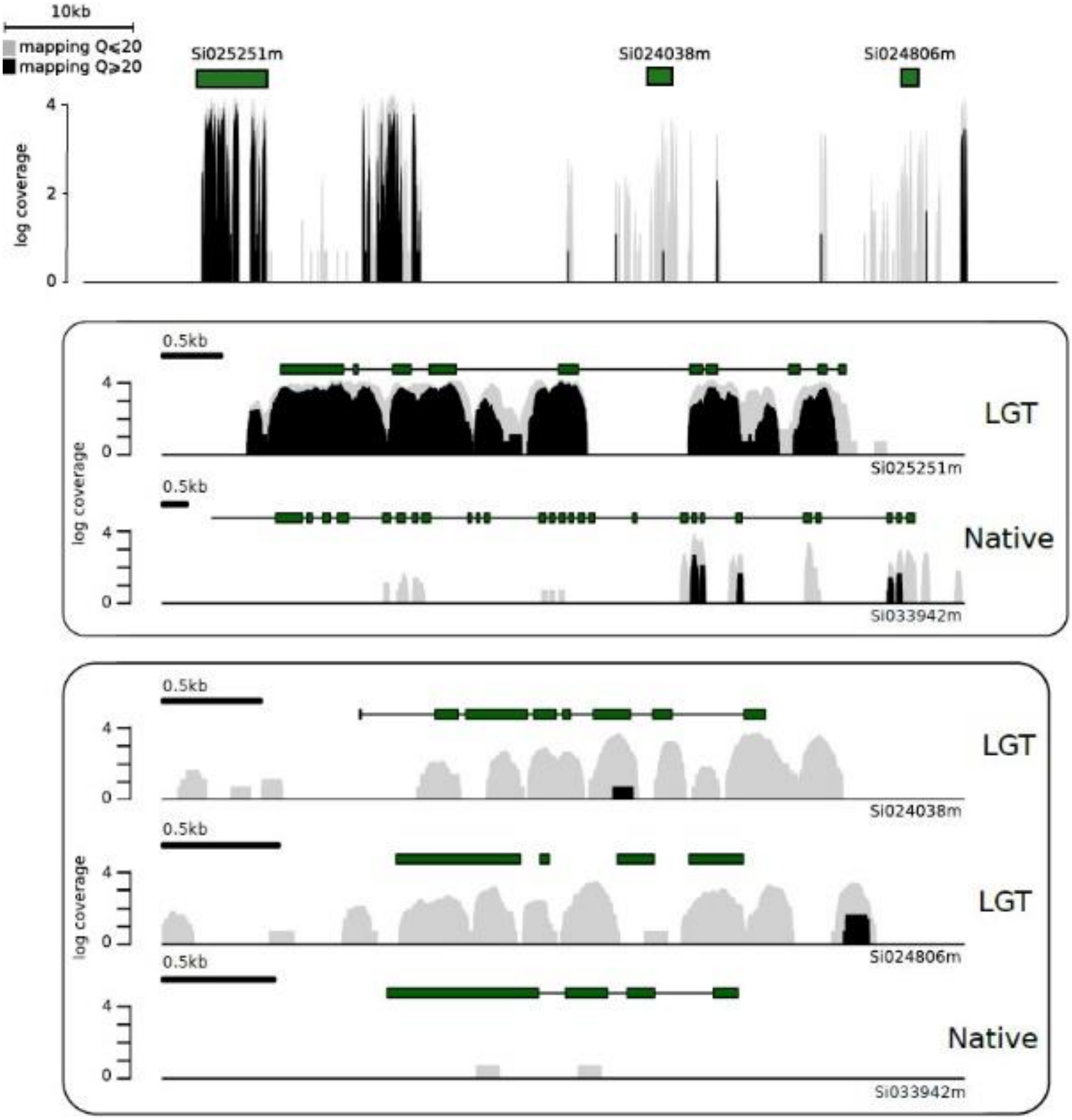


Echinochloa crus−galli SRR5920287
Echinochloa crus−galli SRR5920288
Echinochloa crus−galli SRR5920289
Echinochloa crus−galli SRR5920290
Echinochloa crus−galli SRR5920291
Echinochloa crus−galli SRR5920292
Echinochloa crus−galli SRR5920293
Eragrostis tef SRR1463396
Eragrostis tef SRR1463397
Eragrostis tef SRR1463402
Leersia perrieri SRX663039
Panicum hallii SRR4136572
Frequency
Mean Coverage


Panicum hallii SRR4136573
Panicum hallii SRR4136574
Panicum hallii SRR4136575
Panicum hallii SRR4136576
Panicum hallii SRR5009543
Panicum hallii SRR5009544
Panicum hallii SRR5009545
Panicum virgatum SRR3926366
Panicum virgatum SRR3926427
Panicum virgatum SRR4027878
Setaria italica ERX2710709
Sorghum bicolor ERR2304437
Frequency
Mean Coverage



Sorghum bicolor SRR1945055
Frequency
0 500 1000 1500 2000 2500 3000
0 20 40 60 80 100
Mean Coverage
Sorghum bicolor SRR1946846
Frequency
0 500 1000 1500 2000 2500
0 20 40 60 80
Mean Coverage
Sorghum bicolor SRR5271056
Frequency
0 500 1000 1500 2000
0 10 20 30 40 50 60
Mean Coverage
Triticum aestivum SRR5815659
Frequency
0 500 1000 1500 2000 2500 3000
0 1 2 3 4 5 6
Mean Coverage
Triticum aestivum SRR5893651
Frequency
0 500 1000 1500 2000 2500 3000
0 1 2 3 4 5
Mean Coverage
Zea mays ERR361370
Frequency
0 1000 2000 3000 4000
0 10 20 30
Mean Coverage
Zea mays SRR4039069
Frequency
0 500 1000 1500
0 2 4 6 8 10 12
Mean Coverage
Zea mays SRR5456948
Frequency
0 500 1000 1500
0 1 2 3 4 5 6
Mean Coverage
Zea mays SRR5491191
Frequency
0 500 1000 1500 2000 2500 3000 3500
0 5 10 15 20 25 30 35
Mean Coverage
Zea mays SRR8907142
Frequency
0 1000 2000 3000 4000
0 10 20 30 40
Mean Coverage
Zea mays SRR10128059
Frequency
0 200 400 600 800 1000 1200
0 2 4 6 8 10
Mean Coverage
Zea mays SRR11609573
Frequency
0 1000 2000 3000 4000
0.0 0.5 1.0 1.5 2.0 2.5 3.0
Mean Coverage



Supplementary Figure 2.3: Detail of a laterally acquired fragment in *Setaria italica* genome. In the top panel, the position of three genes laterally acquired from an Andropogoneae species is shown along a portion of chromosome III of *Setaria italica*. The mapping of high-coverage sequence data from the Andropogoneae *Sorghum bicolor* is plotted against the region, with high-quality uniquely mapped reads in black and reads mapped with low quality scores, including those with multiple matches, in grey. All read alignments have a nuclear identity ≥ 90%. The Si024038m and Si024806m loci are recent duplicates, leading to low-quality mapping to them. The bottom panels show expanded schematic of the genes themselves, with exons represented by green boxes. For each LGT, the corresponding native ortholog is also shown.

2.8 Supporting Information

2.8.1 Supplementary Datasets

Supporting datasets available online: https://nph.onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Fnph.17328&file=nph17328-sup-0001-DatasetS1-S4.zip

Supplementary Dataset 2.1: Nucleotide alignments.

Supplementary Dataset 2.2: LGT maximum-likelihood trees, 85 taxa.

Supplementary Dataset 2.3: LGT maximum-likelihood trees, 85 taxa 3rd codon position.

Supplementary Dataset 2.4**:** LGT maximum-likelihood trees, including short-read data.

2.8.2 Supplementary Tables

Supporting table are available online: https://nph.onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1111%2Fnph.17328&file=nph17328-sup-0003-TableS1-S6.zip

Supplementary Table 2.1: List of data sets used in different steps.

Supplementary Table 2.2: Results of the analysis of 17 reference genomes.

Supplementary Table 2.3: Results of the synteny analyses.

Supplementary Table 2.4: Comparisons of the LGT detected in *Alloteropsis semialata* with a previous study

Supplementary Table 2.5: Results of the analysis of LGT among Paniceae species.

Supplementary Table 2.6: Coverage analyses of genes used to compare independently produced sequencing runs.

Chapter 3

3 Repeated lateral transfer of a gene encoding a key C_4 enzyme

Samuel G. S. Hibdige[1], Luke T. Dunning[1], Pascal-Antoine Christin[1]

Keywords: adaptation, evolution, genomics, horizontal gene transfer, lateral gene transfer, phylogenomics, molecular dating, Poaceae, C_4 photosynthesis,

Affiliations: [1] Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom

Personal contributions: I designed the study with the help of my supervisors, performed all the analyses, and wrote the manuscript with the help of my supervisors.

3.1 Abstract

The recent increase in the number of high quality genomes and subsequent comparative studies have led to numerous reports of interspecific gene movements called lateral gene transfers (LGT), some of which have been shown to be a shortcut for biochemical adaptation. One such example is the C_4 photosynthetic gene, encoding the enzyme phosphoenolpyruvate carboxykinase (PCK) that in some grasses has been laterally acquired from distantly related grass species. In this study, we analyse additional grasses to look for further evidence of the relative contributions of LGT to the origins of C_4-specific PCK. We reconstruct the phylogeny of genes encoding PCK, and show that *pck* genes in multiple *Echinochloa* species were also acquired via LGT from Cenchrinae, mirroring the pattern observed in the genus *Alloteropsis*, Even though *Alloteropsis* and *Echinochloa* are closely related, phylogenetic trees support independent LGT into each of these genera. Furthermore, reanalyses of *pck* genes from Chloridoideae grasses reveal multiple LGT of the genes encoding the C_4-specific forms of the enzyme within this subfamily. These results indicate that genes for PCK have been moved across distant grass species multiple times independently. The C_4-specific isoform of PCK improves the efficiency of the C_4 pathway in some conditions, but the non- C_4 PCK function ancestrally encoded by a single gene must be retained. Lateral gene transfer therefore provides an alternative to gene duplication followed by neofunctionalization, making it beneficial in multiple C_4 lineages.

3.2 Introduction

The evolution of novel adaptations is driven by selection acting on the genetic variation existing within populations. Mutations happen constantly, and when disadvantageous are rapidly removed by purifying selection. The mutations that persist and are subject to selection are those that are neutral, evolving strictly under genetic drift, and those that are advantageous and gradually increasing in frequency. Such mutations can consist of substitutions or small indels in specific genes, but also larger genomic rearrangements, including those that suppress or duplicate genomic fragments. Among those, duplications of genes are thought to play an important role in evolutionary transitions, as the duplicates generate genetic redundancy, so that mutations of one of the copies are less likely to be deleterious (Zhang 2003). One of the duplicates can therefore come to fulfil novel functions, in a process called neofunctionalization, which has been associated with important innovations in a variety of organisms (Zhang 2006; Deng *et al*., 2010; Lyu *et al*., 2020). In classical models, the duplicates originate in a single genome, through some DNA copy mistake. The impact might however be similar if slightly divergent copies of homologous genes come to exist in a genome following genetic exchanges among organisms, a process happening across the whole genome following allopolyploidization (Wang *et al*., 2006; Ha *et al*., 2009). The recent advent of genomic analyses has revealed multiple cases of interspecific gene movements, called lateral gene transfers (LGT; Dunning *et al*., 2019). Some of the transfers have been shown to shortcut biochemical adaptation (Phansopa *et al*., 2020), but whether they circumvent the need for gene duplication and neofunctionalization remains to be assessed.

C_4 photosynthesis is a complex trait that results from the coordinated action of multiple enzymes in specific leaf compartments to boost productivity in tropical conditions (Hatch, 1987; Sage, 2004). All enzymes of the C_4 pathway existed in the non- C_4 ancestors, but they were responsible for different functions (Aubry *et al*., 2013). During the assembly of C_4 photosynthesis, these multiple enzymes drastically increased in abundance and underwent alterations of their kinetics and spatial localisation (Svensson *et al*., 2003; Alvarez *et al*., 2019), in a typical pattern of neofunctionalization. C_4 photosynthesis is therefore an outstanding system to evaluate the genomic events that facilitate gene co-option into novel functions, and its origin was originally hypothesised to be a textbook example of gene duplications driving functional innovation (Monson *et al*., 2003). Subsequent analyses of genomes however failed to find evidence for consistent duplication of genes encoding C_4 enzymes (Williams *et al*., 2012), and phylogenetic analyses of specific C_4-related genes have found that many were co-opted for the fundamentally different C_4 function without prior duplication, allowing the persistence of copies responsible for the ancestral enzyme function (Christin *et al*., 2007, 2009b). Among the genes encoding C_4 enzymes analysed so far, only those for phosphoenolpyruvate carboxykinase (PCK) appear to have been consistently duplicated recently before their co-option for C_4, in each case retaining a copy that presumably maintained the ancestral function (Christin *et al*., 2009a). In one group of species, the genes for PCK used in the C_4 pathway were however laterally-acquired from distantly related species, and cohabit in the recipient genome with ancestral homologs not involved in C_4 photosynthesis (Christin *et al*., 2012a; Dunning *et al*., 2019). The events leading to C_4-specific *pck* genes in a few other groups of grasses were ambiguous (Christin *et al*., 2009a), and their history now need to be reanalysed with the newly available genomic information to assess the relative contributions of gene duplications and lateral gene transfers to the origins of C_4-specific PCK.

In this study, we reconstruct the phylogeny of genes encoding PCK based on data extracted from available genomes. We specifically test whether the unexpected patterns previously reported for some C_4 grasses can result from lateral gene transfers. Besides the laterally acquired genes in *Alloteropsis semialata* (Christin *et al*., 2012a), those of *Echinochloa* species presented affinities with the same group of potential donors (Moreno-Villena *et al*., 2018). Those from the large C_4 group called Chloridoideae present patterns originally interpreted as the fingerprint of gene duplications followed by differential gene losses (Christin *et al*., 2009a), but which might be seen as LGT in the light of recent reports of interspecific gene transfers. For each of these groups, we combine stringent phylogenetic tests and molecular dating to carefully evaluate the likelihood of lateral transfer leading to the acquisition of their C_4-specific PCK.

3.3 Materials and Methods

3.3.1 Species sampling

A database of the available grass genomes and transcriptomes was compiled (Supplementary Table 3.1). Where genome annotation was available, the coding sequences were blasted using the *pck* gene Alloteropsis_semialata_FR845830.1 as a reference and blast hits with a length of over 400 bp and an E value of $\leq 1e^{-20}$ were considered candidates. Where annotation was not available, the assembly itself was blasted using Alloteropsis_semialata_FR84583.1 as a reference and where there were blast hits with $\leq 1e^{-20}$ the coordinates of the hit and 1,000 bp flanking regions either side were extracted. In the case that only unassembled reads were available, the reads were blasted using Blastn 2.2.31+ and assembled in Geneious Prime 2019 using the default settings. Finally the reference *pck* gene was blasted on NCBI for additional sequences. Additional *pck* genes from *Musa acuminata* and *Ananas comosus* from Ensembl plants were included as an outgroup.

3.3.2 Phylogenetic analyses

An alignment of the raw sequences was generated in Geneious using Muscle (default parameters) and manually edited. Flanking regions and UTRs were removed so that only introns and exons remained and sequences less than 1,000 bp were deleted. The coding sequence alignment was created by removing the introns from the alignment. The alignment was translated and manually checked to ensure the correct reading frame. The 3rd position of each codon was extracted to generate the 3rd codon position alignment. The intron alignment was generated from the originally genomic alignment, transcriptome data and sequences lacking introns were removed. The alignment of the introns was manually edited and the exons removed.

For each dataset, the best-fit substitution model was identified using the SMS algorithm (Lefort, Longueville and Gascuel, 2017). Phylogenetic trees were then generated for each partition using PhyML (Guindon *et al.*, 2010) to infer maximum likelihood trees with 100 bootstrap replicates. Bayesian trees were generated for each partition using MrBayes with the parameters 8 Chains, 10,000,000 generations and 2 runs. The runs were monitored using Tracer (Rambaut *et al.*, 2018) and the burn in period was set to 10,000,000. A majority consensus was then inferred from the posterior trees.

3.3.3 Molecular dating

Molecular dating was used to estimate the relative ages of the two groups of Chloridoideae, specifically testing the hypothesis that their ages differ, which would not be expected in the case of a deep gene duplication. A subset of the dataset was taken to represent each group where the same species was represented in each data set.

Divergence times were estimated using BEAST version v2.6.2 with two independent MCMC tree runs (100,000,000 generations, sampling every 1,000, burn in period of 10,000,000 generations, GTR +G substitution model, Log-normal relaxed clock, Yule process speciation prior, root calibrated with normal distribution, with a mean of 51.2 mya (million years ago) and a standard deviation of 0.001 (based on estimate from Christin *et al.*, 2014). Tracer was used to examine the convergence of the runs.

3.4 Results

3.4.1 Phylogenetic patterns suggest two transfers within Paniceae

The phylogenetic tree inferred from full coding sequences of genes encoding PCK (*pck*) globally matches the expected species tree for the grass family, with a monophyletic PACMAD clade sister to Pooideae and then Oryzoideae, and monophyletic Chloridoideae and Panicoideae subfamilies within the PACMAD clade (Figure 3.1). Within Panicoideae, the Andropogoneae and Paspaleae tribes are sister to Paniceae, as expected. Multiple gene duplications specific to some sublineages of Paniceae are apparent, including in *Digitaria*, Melinidineae, *Cenchrus*, *Paspalum*, and *Zea* (Figure 3.1). Four of these gene duplications were detected before and preceded the co-option of *pck* genes into the C_4 pathway of these groups (Christin *et al.*, 2009a). The new analyses, based on vastly increased amounts of data, therefore confirm that *pck* genes have generally been duplicated before one of the copies was recruited into the new C_4 function.

Despite the overall congruence with the species tree, important deviations from the species tree are observed within Paniceae. As previously reported (Christin *et al*., 2012a; Dunning *et al*., 2017), two groups of *pck* genes are retrieved from the *Alloteropsis* genus. A first group named *pck-1P1* (Dunning *et al*., 2017) is detected in all three *Alloteropsis* species and is placed as expected based on the species tree, sister to *Entolasia marginata* (Figure 3.1). A second group, detected only in *A.angusta* and *A. semialata*, is nested with bootstrap support within Cenchrinae. This gene, named *pck-1P1_LGT-C*, is the one used for the C_4 pathway of these two *Alloteropsis* species (Dunning *et al*., 2017). Two groups of *pck* genes are similarly detected in the *Echinochloa* species, and both are represented by multiple species (Figure 3.1). While the first one is placed outside of the other subtribes as expected based on the species tree, the second one is also nested within Cenchrinae, and is placed as sister to genes from *Cenchrus* with high statistical support (Figure 3.1). This result confirms that *pck* genes from *Echinochloa* were likely acquired via LGT from Cenchrinae (Dunning *et al*., 2019), but the wider sampling of *Echinochloa* species achieved here reveals that the LGT *pck* is present in numerous species from the genus (Figure 3.1). Importantly, the *Echinochloa* and *Alloteropsis* LGT, despite both coming from Cenchrinae, do not group together (Figure 3.1). These patterns suggest that Cenchrinae transferred *pck* genes independently to each of *Echinochloa* and *Alloteropsis*.

The inferred relationships remained similar when considering only the 3rd positions of codons, which are less subject to selection. In particular, some *Alloteropsis* and *Echinochloa pck* genes are still nested within Cenchrinae, with *Echinochloa* sister to a group of *Cenchrus* sequences and *Alloteropsis* then sister to both of them (Figure 3.2). The use of 3rd positions of codons rules out adaptive evolution as a bias explaining the results, and the patterns thus unequivocally support two transfers of *pck* genes from Cenchrinae.

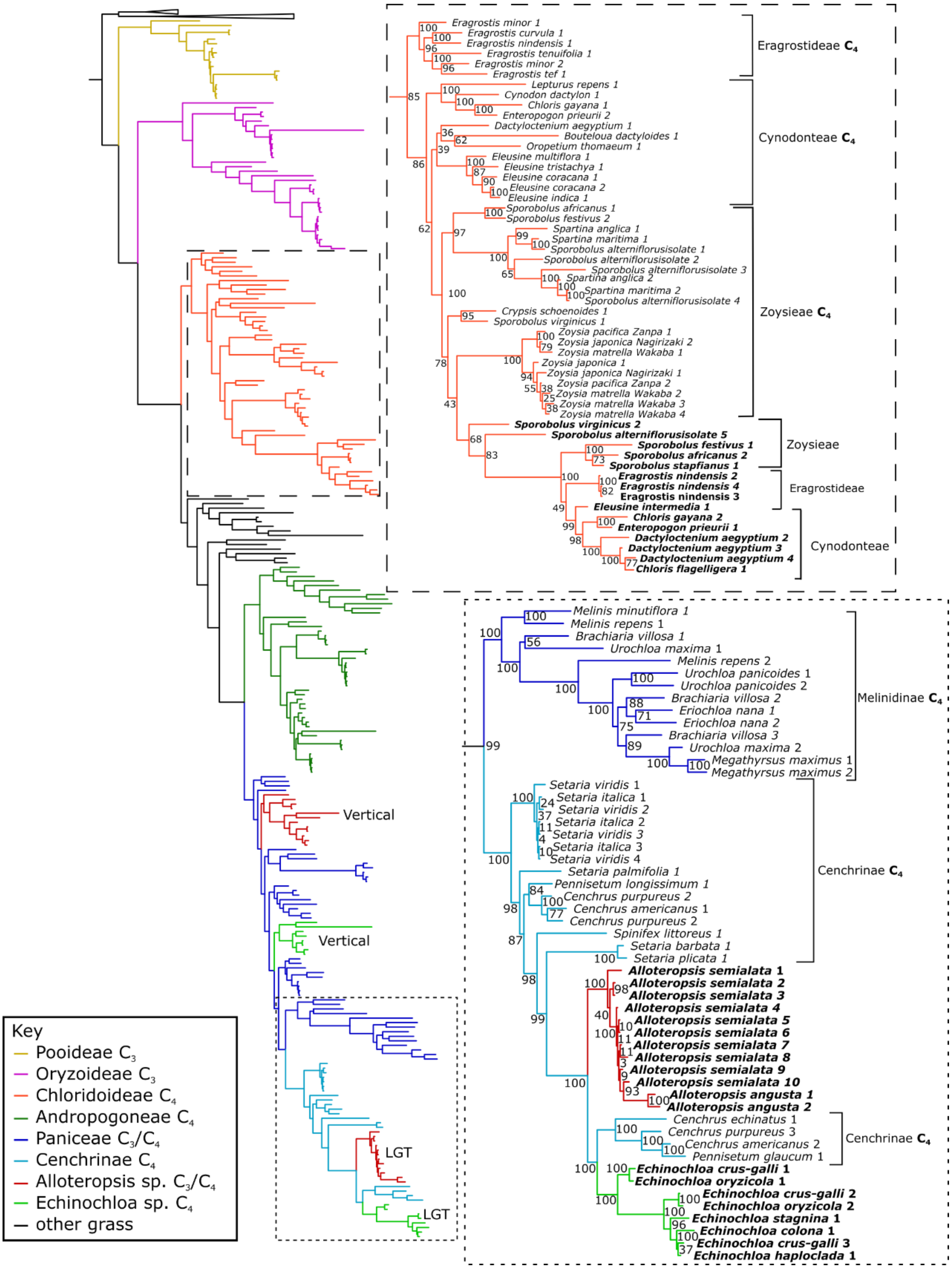


Figure 3.1: Consensus Bayesian tree of the *pck* gene CDS in grasses, generated with MrBayes. Colours denote different clades and node numbers represent the probability as a percentage. Genes in bold indicate laterally acquired *pck* genes. The tribes indicated are described in Soreng *et al*., 2015 and their photosynthetic type is listed as C3 or C4. Numbers next to species relate to the accession ID found in Supplementary Table 3.1

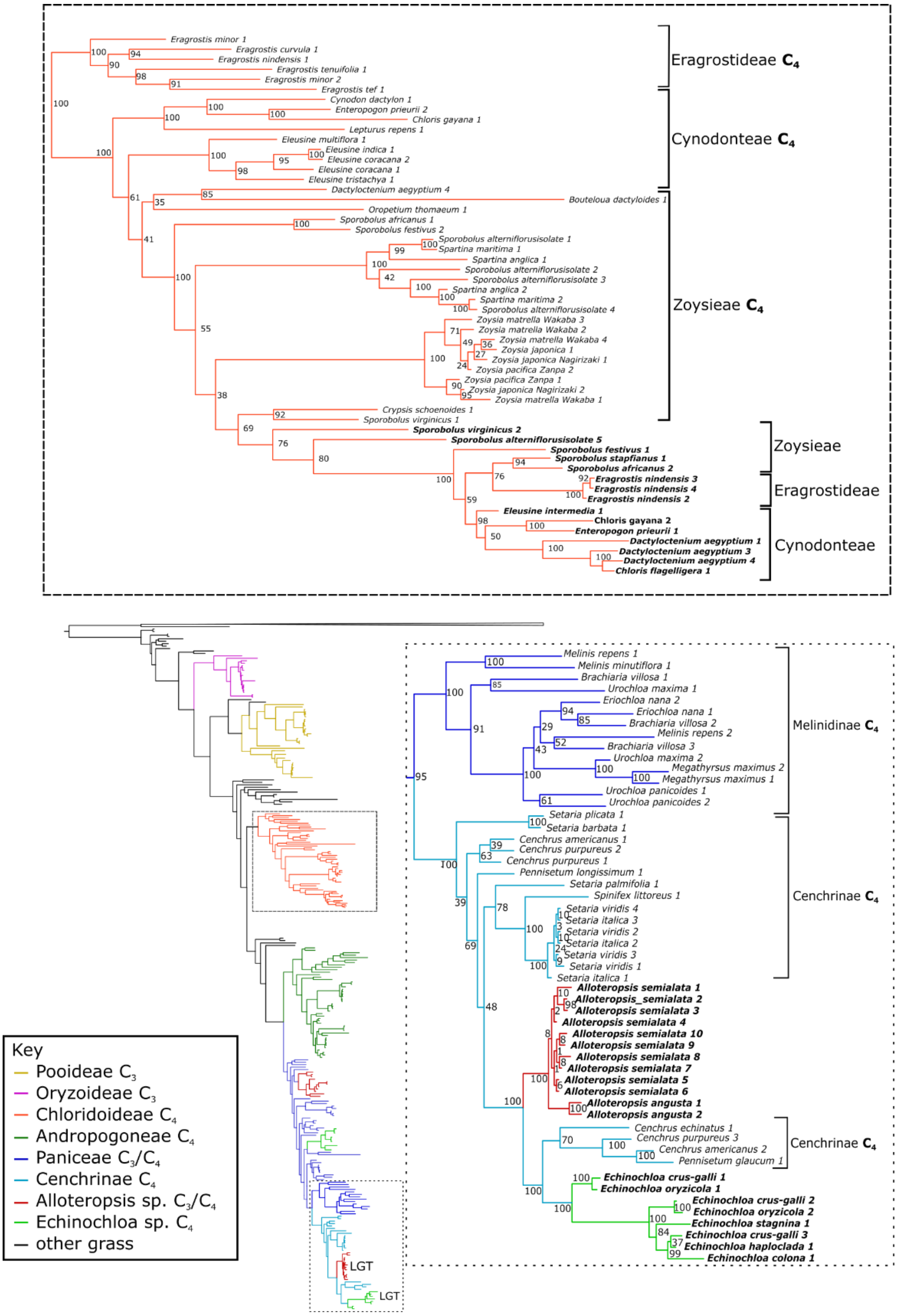


Figure 3.2: Consensus Bayesian tree of the *pck* gene 3rd Codon positions in grasses, generated with MrBayes. Colours denote different clades and node numbers represent the probability as a percentage. Genes in bold indicate laterally acquired *pck* genes. The tribes indicated are described in Soreng *et al*., 2015 and their photosynthetic type is listed as C3 or C4. Numbers next to species relate to the accession ID found in Supplementary Table 3.1

3.4.2 Multiple transfers of pck genes among Chloridoideae

Outside of Panicoideae, discrepancies with the species tree are also observed within the large, wholly C_4 Chloridoideae subfamily (Figure 3.1). *Eragrostis* sequences are sister to the rest of the family, as expected based on the species tree. The diversity of the subfamily is then represented in a pattern consistent with the species tree, with a gene duplication leading to the C_4-specific genes from the *Spartina* group (Figure 3.1, named *pck-B* by Christin *et al*., 2009a). However, a group of highly similar sequences capturing again the diversity of the subfamily is nested within those of the *Sporobolus* genus (Figure 3.1). This group of genes, which contains genes used for the C_4 pathway, was previously named *pck-B* and based on a more superficial sampling to correspond to a duplicate originating at the base of the subfamily (Christin *et al*., 2009a). The patterns revealed here with a large species sampling are well supported, even when considering only the 3rd positions of codons, and are highly incompatible with the known species relationships, urging for a re-evaluation of the deep duplicate hypothesis.

Under a deep duplication scenario, the two copies found within some species would have diverged at the same time, which should predate the age of the subfamily. To test this hypothesis, we estimated the crown of the Chloridoideae, independently for each of the two sets of *pck* genes. In each case, we included the same set of species outside of Chloridoideae. Within Chloridoideae, the same set of species was used, but we first considered those genes not part of the group nested within *Sporobolus* and we then considered only those genes nested within *Sporobolus*. The two analyses yielded similar ages for groups outside of Chloridoideae (Figure 3.3a, and Figure 3.3b), confirming that they represent a fair assessment of the relative ages of groups of grasses. In stark contrast, massively different ages were estimated for the two groups of *pck* genes from Chloridoideae (Figure 3.3c). The crown age of those not nested within Chloridoideae was estimated around 35 Ma (median = 34.71 Ma, confidence interval = 24.16-45.58 Ma), an age compatible with those obtained for the subfamily with other markers (Christin *et al*., 2008). By contrast, the most recent common ancestor (mrca) age of the group nested within *Sporobolus* was estimated at 17 Ma (median = 17.55 Ma, confidence interval = 9.72-27.17 Ma), even in the absence of other Chloridoideae sequences. These analyses confirm that sequences from this group diverged after the species that bear them, a pattern that can be explained by movements of genes among established species. We conclude that *pck* genes were laterally transferred among Chloridoideae species, and the phylogenetic patterns suggest this process happened multiple times independently.

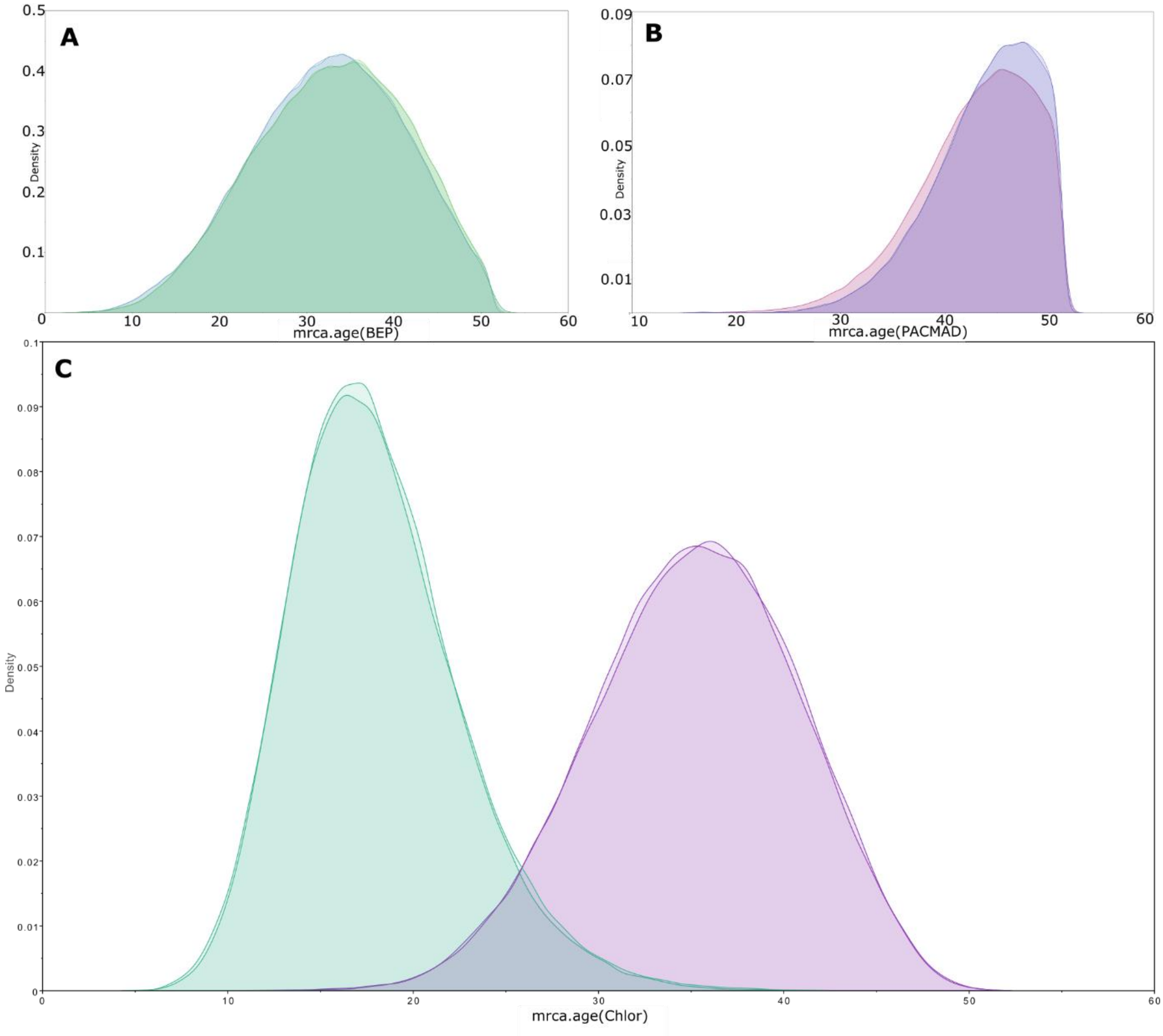


Figure 3.3: Distribution of posterior age estimations for the most recent common ancestor (mrca), using the two sets of *pck* genes found within Chloridoideae. A shows an estimate for the BEP grasses, B shows an age estimate for the PACMAD grasses, C shows an age discrepancy for the mrca between the two sets of Chloridoideae *pck* genes suggesting that the sequences from one group diverged after the species that bear them. In each case, posterior distributions are shown for two independent analyses.

3.5 Discussion

3.5.1 Multiple lateral gene transfers in Panicoideae and in Chloridoideae

The phylogenetic patterns reported here confirm placements of genes encoding phosphoenolpyruvate carboxykinase (PCK; *pck* genes) incompatible with the species tree for two genera within the Paniceae tribe; *Alloteropsis* and *Echinochloa* (Figure 3.1). The nesting of *Alloteropsis pck* genes within those of Cenchrinae was noted before, first based on Sanger sequences (Christin *et al.*, 2012a) and then using transcriptome and genome data (Dunning *et al.*, 2017, 2019). After careful analyses of non-coding flanking regions, these phylogenetic patterns led to one of the first conclusive cases of plant-to-plant transfer of nuclear protein-coding genes (Christin *et al.*, 2012a). In a subsequent analysis, it was observed that genes from the *Echinochloa* genus also group with those of Cenchrinae (Dunning *et al.*, 2019), a conclusion confirmed here with a broader species sampling (Figure 3.1). Importantly, this Cenchrinae *pck* gene was detected in all *Echinochloa* species we screened here, suggesting it was acquired laterally early during the diversification of *Echinochloa*. While *Echinochloa* and *Alloteropsis* both received *pck* genes from Cenchrinae, sequences from these two genera do not form a monophyletic group, as *Echinochloa* is closer to the sequences of some sampled Cenchrinae (Figure 3.1). This phylogenetic pattern indicates that the same group of Paniceae (Cenchrinae) provided *pck* genes independently to *Alloteropsis* and *Echinochloa*.

Besides these two transfers of *pck* genes among Paniceae, our analyses also suggest that *pck* genes have been moved among distinct Chloridoideae species. Indeed, the detection of two distinct *pck* genes in a number of distantly-related Chloridoideae would be compatible with a duplication in the early history of the group, but our dating analyses indicates that one of them diversified long after the species that possess it. Such a pattern can be explained by genetic exchanges among reproductively isolated species. Because the relationships among species based on one of the *pck* genes strongly differ from the species tree of Chloridoideae, under an LGT scenario the genes must have been passed multiple times among species of this subfamily. The gene phylogenetic tree indicates that the gene originated in the *Sporobolus* genus, and it was subsequently moved into species belonging to four different clades; *Eragrostis*, *Eleusine*, *Chloris*/*Enteropogon*, and *Dactyloctenium* (Figure 3.1).

Although LGT within the Chloridoideae is supported, there are alternative scenarios that could explain the results. Phylogenetic trees can be influenced by a number of factors that cause gene trees to appear discordant with the species tree, as discussed in Chapter 2. The nesting within the Chloridoideae does not follow a pattern congruent with gene duplication at the base of the clade but could instead be due to gene fusion or assembly errors. Gene fusion will cause inconsistencies in phylogenetic trees as the constituent elements will have potentially had different evolutionary histories (Yanai, Wolf & Koonin, 2002). Assembly errors may cause an area to be presented as a resolved region but in reality is a chimaera of two similar areas as you might expect in paralogs. Phylogenetic trees containing such sequences will be incongruent with the species tree.

Molecular dating relies on the turnover of genetic material to estimate divergence time. Duplicates existing in different parts of the genome may be subject to different selection pressures and therefore rates of turn over (Som, 2015). Paralogs often exist in different parts of the genome as a result of the duplication process or transposable elements. Selection pressures can vary wildly between paralogs post duplication causing rapid changes between the two as seen in neofunctionalization and subfunctionalisation (Rastogi and Liberles, 2005). Under a neofunctionalization scenario, a gene exist under purifying selection pre-duplication. Post duplication, one paralog may remain under purifying selection whilst the second may experience relaxed or directional selection towards a new function (Wagner, 2002). When compared via molecular dating the paralog under purifying selection will appear younger due to lower relative rate of change. To rule this out an analysis of rates could be carried out on the two groups of paralogs.

Among the subfamilies Panicoideae and Chloridoideae, our analyses identified patterns suggestive of at least six independent lateral gene transfers of *pck* genes, in several cases from the same source. A transcriptome analysis from a previous species (*Cymbopogon*, in Andropogoneae) further suggested a seventh transfer, although it could not be supported with high confidence (Dunning *et al.*, 2019). For most grass species, *pck* sequences are not available, so that the total number of lateral gene transfers of this gene might be even higher. Our analyses already indicates that *pck* genes were recurrently moved among distinct grass species, suggesting that this gene is especially prone to such movements.

3.5.2 Requirement for gene duplications favoured lateral gene transfers

The large numbers of lateral gene transfers of *pck* genes might be linked to the propensity of the gene to physically move among grasses, but we favour the hypothesis that the post-transfer retention of the gene is more likely than for most other genes. The encoded enzyme plays a key role in the C_4 pathway of some species, but it always acts in combination with other enzymes that play a similar function (Prendergast *et al.*, 1987; Kanai and Edwards, 1999; Wang *et al.*, 2014). In many large C_4 lineages, only a subset of species use the PCK enzyme, strongly suggesting that this enzyme was added after the origin of the C_4 trait, during its subsequent evolutionary improvements (Christin and Osborne, 2014). Indeed, the addition of the biochemical shuttle based on PCK increases the range of light conditions in which the plants are likely to be competitive (Bellasio and Griffiths, 2014), likely allowing transitions to new habitats. While the acquisition of the PCK shuttle was likely advantageous, it might have been complicated by the low number of *pck* copies existing in plant genomes.

The *pck* gene family ancestrally consists of a single copy (Shen *et al.*, 2017; Moreno-Villena *et al.*, 2018), which encodes a protein responsible for various non-photosynthetic functions in non-C_4 plants (Leegood and Walker, 2003; Shen *et al.*, 2017). The evolution of C_4-specific PCK involved increases of enzymatic activity in the leaf (Shen *et al.*, 2017), important upregulation of the genes specifically in the photosynthetic leaves (Moreno-Villena *et al.*, 2018), and positive selection on *pck* coding sequences that likely altered the enzyme catalytic properties (Christin *et al.*, 2009a; Moreno-Villena *et al.*, 2018). However, in all C_4 grasses analysed so far, the C_4-specific *pck* co-exist with a *pck* copy expressed at low levels and without evidence of past positive selection, this copy likely retains the ancestral function (Figure 3.1; Christin *et al.*, 2009a; Moreno-Villena *et al.*, 2018). This implies gene duplication directly preceded the co-option for C_4 photosynthesis, and that duplicates of *pck* that are not co-opted for C_4 photosynthesis do not persist over long evolutionary times as duplicates are rare in C_3 grasses. There is therefore a limited number of opportunities and a short window of time for plants to co-opt *pck* genes into their C_4 pathway, limiting the evolvability of a PCK shuttle.

All the plants that laterally acquired a *pck* gene are C_4 species, as are all the donors. Because these plants were already performing C_4 photosynthesis before the transfers (Christin *et al.*, 2009a; Dunning *et al.*, 2017), the acquisition of the foreign gene directly led to a PCK shuttle with its associated advantages. Importantly, the requirement for a gene duplication followed by changes in the expression pattern and coding sequences for the gene means that many C_4 species in which a PCK would be beneficial might not be able to easily evolve using their native genetic material. Lateral transfers of *pck* genes are thus likely to be advantageous, so that positive selection rapidly leads to the spread of fixation of foreign *pck* in the recipient species. We therefore propose that the high frequency of lateral transfers of *pck* revealed here results from the difficulty of evolving C_4 *pck* from native genes together with the advantages of the encoded trait for the numerous C_4 plants lacking a PCK-based pathway.

3.6 Conclusions

In this study, we revisit the history of genes encoding the key C_4 enzyme phosphoenolpyruvate carboxykinase (PCK) in grasses. We confirm that C_4-specific genes evolved multiple times following the duplication of non-C_4 genes, but also that some C_4-specific genes for PCK were laterally transferred twice to some C_4 Paniceae. A careful reanalysis of the large Chloridoideae subfamily further suggests a minimum of four lateral gene transfers within the group, leading to at least six independent transfers of this gene. We suggest that the requirement for gene duplication limits the ability of C_4 plants to add a PCK shuttle that would boost their C_4 pathway. Lateral transfers of genes for PCK therefore provide a direct advantage to plants lacking such shuttle, so that they are rapidly fixed by positive selection. We conclude that lateral gene transfers offer an alternative to gene duplication followed by neofunctionalization in some groups of plants.

3.7 Acknowledgements

This work was funded by a Royal Society Research grant (RGF\EA\180247). P.-A.C. is supported by a Royal Society University Research Fellowship (URF\R\180022). L.T.D is supported by a Natural Environment Research Council Independent Research Fellowship (NE/T011025/1).

3.8 Supplementary Tables

Supplementary Table 3.1 Genetic resources, Ploidy level estimates derived from Chromosome Counts Database (Rice et al, 2015).

Species	Ploidy Level	Resource	GENE/ACCESSION
Acroceras tonkinense	NA	NCBI	FM211817
Acroceras zizanioides	4x	BioProject PRJNA395007	AZIZ_c31202_g1_i1
Aegilops tauschii	2x	Ensembl Plants	AET4Gv20558700
Aegilops tauschii subsp. tauschii	2x	NCBI	XM_020320501
Alloteropsis angusta	2x	NCBI	FR845842
Alloteropsis angusta	2x	NCBI	FR845843
Alloteropsis angusta	2x	NCBI	FR845844
Alloteropsis angusta 1	2x	NCBI	FR845845
Alloteropsis angusta 2	2x	NCBI	FR845846
Alloteropsis angusta	2x	NCBI	KX788100
Alloteropsis cimicina	2x	NCBI	FR845848
Alloteropsis cimicina	2x	NCBI	FR845849
Alloteropsis cimicina	2x	NCBI	FR845850
Alloteropsis cimicina	2x	NCBI	KX788094
Alloteropsis semialata	2x, 6x, 8x 12x	NCBI	KX788088
Alloteropsis semialata	2x, 6x, 8x 12x	NCBI	KX788089
Alloteropsis semialata	2x, 6x, 8x 12x	NCBI	KX788090
Alloteropsis semialata	2x, 6x, 8x 12x	NCBI	KX788091

Species	Ploidy Level	Resource	GENE/ACCESSION
Alloteropsis semialata	2x, 6x, 8x 12x	NCBI	KX788092
Alloteropsis semialata	2x, 6x, 8x 12x	NCBI	KX788093
Alloteropsis semialata	2x, 6x, 8x 12x	NCBI	KX788095
Alloteropsis semialata	2x, 6x, 8x 12x	NCBI	KX788096
Alloteropsis semialata	2x, 6x, 8x 12x	NCBI	KX788097
Alloteropsis semialata	2x, 6x, 8x 12x	NCBI	KX788098
Alloteropsis semialata	2x, 6x, 8x 12x	NCBI	KX788099
Alloteropsis semialata	2x, 6x, 8x 12x	NCBI	KX788101
Alloteropsis semialata	2x, 6x, 8x 12x	NCBI	KX788102
Alloteropsis semialata	2x, 6x, 8x 12x	NCBI	KX788103
Alloteropsis semialata	2x, 6x, 8x 12x	NCBI	KX788104
Alloteropsis semialata	2x, 6x, 8x 12x	NCBI	KX788105
Alloteropsis semialata	2x, 6x, 8x 12x	NCBI	KX788106
Alloteropsis semialata	2x, 6x, 8x 12x	NCBI	KX788107
Alloteropsis semialata	2x, 6x, 8x 12x	NCBI	KX788108
Alloteropsis semialata	2x, 6x, 8x 12x	NCBI	KX788109
Alloteropsis semialata subsp. eckloniana	2x, 6x	NCBI	FR845829
Alloteropsis semialata subsp. semialata	2x	NCBI	FR845830
Alloteropsis semialata subsp. semialata 5	2x	NCBI	FR845831
Alloteropsis semialata subsp. semialata	2x	NCBI	FR845832
Alloteropsis semialata subsp. semialata	2x	NCBI	FR845833

Species	Ploidy Level	Resource	GENE/ACCESSION
Alloteropsis semialata subsp. semialata	2x	NCBI	FR845834
Alloteropsis semialata subsp. Semialata 8	2x	NCBI	FR845836
Alloteropsis semialata subsp. semialata	2x	NCBI	FR845837
Alloteropsis semialata subsp. Semialata 6	2x	NCBI	FR845840
Alloteropsis semialata subsp. Semialata 9	2x	NCBI	FR845841
Alloteropsis semialata subsp. Semialata 7	2x	NCBI	FR845986
Alloteropsis semilata_angusta	NA	NCBI	AANG_SEQUENCES
Alloteropsis semilata_AUS1 4	2x	NCBI	AUS1_17510
Alloteropsis semilata_KWT	2x	NCBI	KWT3_07097
Alloteropsis semilata_KWT	2x	NCBI	KWT3_07097
Alloteropsis semilata_LO4 1	2x	NCBI	L04B_01368
Alloteropsis semilata_LO4	2x	NCBI	L04B_32147
Alloteropsis semilata_ZAM 2	2x	NCBI	ZAM15-05-10_43371
Alloteropsis semilata_ZAM 1	2x	NCBI	ZAM15-05-10_43373
Alloteropsis semilata_ZAM	2x	NCBI	ZAM15-05-10_59661
Alloteropsis_angusta_MRL	2x	NCBI	MRL48_032374
Alloteropsis_angusta_MRL	2x	NCBI	MRL48_004156
Alloteropsis_semialata_AUS1	2x	NCBI	AUS1_05378
Ananas comosus	2x, 3x, 4x	Ensembl Plants	Contig6:678453-6678728
Arabidopsis thaliana	2x	NCBI	PCK1
Arabidopsis thaliana	2x	NCBI	PCK2

Species	Ploidy Level	Resource	GENE/ACCESSION
Arabidopsis thaliana	2x	NCBI	AT5G65690.1
Arabidopsis thaliana	2x	NCBI	AT4G37870.1
Aristida rhiniochloa	2x	NCBI	FM211819
Arundinaria sp. Hodkinson s.n.	NA	NCBI	FM211820
Arundinaria sp. Hodkinson s.n.	NA	NCBI	FM211821
Arundo donax	36-54	NCBI	FM211822
Austroderia richardii	NA	NCBI	FM211833
Bonia amplexicaulis	NA	http://www.genobank.org/bamboo	scaffold209:80864:86094
Bonia amplexicaulis	NA	http://www.genobank.org/bamboo	scaffold5224:175284:179608
Bonia amplexicaulis	NA	http://www.genobank.org/bamboo	scaffold3415:121032:125384
Bouteloua dactyloides 1	NA	BioProject PRJNA395007	GARE_GARE01026912.1
Brachypodium distachyon	2x	Phytozome	Bradi1g67730.6
Brachypodium distachyon	2x	Ensembl Plants	BRADI_1g67730v3
Brachypodium distachyon	2x	NCBI	XM_003558272
Brachypodium distachyon	2x	NCBI	XM_010230554
Brachypodium distachyon	2x	NCBI	Bradi1g67730.1
Brachypodium stacei	NA	Phytozome	Brast02G119500.1
Brachypodium hybridum	NA	Phytozome	Brahy.S02G0125700
Brachypodium hybridum	NA	Phytozome	Brahy.D01G0929700
Brachypodium mexicanum	4x	Phytozome	Brame.02UG321700.1
Brachypodium mexicanum	4x	Phytozome	Brame.02PG144800.1

Species	Ploidy Level	Resource	GENE/ACCESSION
Brachypodium sylvaticum	2x, 6x	Phytozome	Brasy2G145200.1
Bromus hordeaceus	2x, 4x	NCBI	FM211826
Cenchrus americanus	2x	NCBI	FR872788
Cenchrus americanus 1	2x	NCBI	MK167362
Cenchrus americanus 2	2x	BioProject PRJNA395007	GEUY_GEUY01000157.1
Cenchrus echinatus 1	2x, 4x	NCBI	FM211827
Cenchrus longissimus	NA	NCBI	FM211867
Cenchrus purpureus	4x	BioProject PRJNA395007	GWHAORA00000000
Centropodia forskaolii	NA	NCBI	FM211828
Chandrasekharania keralensis	NA	NCBI	MK737796
Chasmanthium latifolium	NA	NCBI	FM211829
Chasmanthium latifolium	NA	BioProject PRJNA395007	CLAT_c22086_g1_i1
Chionochloa macra	6x	TSA database	GFMB_GFMB01232765.1
Chionochloa pallens	6x	TSA database	GHUI_GHUI01107594.1
Chloris flagellifera 1	NA	BioProject PRJNA395007	GGLS_GGLS01036505.1
Chloris gayana 1	2x, 3x, 4x	NCBI	FM211830
Chloris gayana 2	2x, 3x, 4x	NCBI	FM211831
Coix lacryma-jobi	2x, 4x	NCBI	FM211832
Coix lacryma-jobi var. ma-yuen	2x, 3x	http://www.phyzen.com/adlay/	Adlay_V1-2_transcripts.fasta
Coix aquatica	NA	NCBI	EVM0029670
Cucumis sativus	2x	NCBI	L31899.Cucumber

Species	Ploidy Level	Resource	GENE/ACCESSION
Cynodon dactylon 1	2x, 9x	NCBI	FM211835
Cyrtococcum patens	2x, 4x	NCBI	FM211883
Cyrtococcum patens	2x, 4x	BioProject PRJNA395007	CPAT_c42379_g1_i1
Dactyloctenium aegyptium 1	2x, 4x	NCBI	FM211836
Dactyloctenium aegyptium 2	2x, 4x	NCBI	FM211837
Dactyloctenium aegyptium 3	2x, 4x	BioProject PRJNA395007	DAEG_c22722_g1_i1
Dactyloctenium aegyptium 4	2x, 4x	BioProject PRJNA395007	DAEG_c22722_g1_i2
Danthonia californica	4x	BioProject PRJNA395007	DCAL_c23007_g1_i1
Dichanthelium cumbucana	NA	NCBI	FR872787
Digitaria ciliaris	2x, 4x, 6x, 8x	BioProject PRJNA395007	DCIL_c30956_g1_i2
Digitaria ciliaris	2x, 4x, 6x, 8x	BioProject PRJNA395007	DCIL_c30956_g1_i1
Digitaria didactyla	2x, 4x, 8x	NCBI	FM211838
Digitaria didactyla	2x, 4x, 8x	NCBI	FM211839
Digitaria sanguinalis	2x, 4x, 6x, 8x	NCBI	FM211840
Digitaria sanguinalis	2x, 4x, 6x, 8x	NCBI	FM211841
Digitaria sanguinalis	2x, 4x, 6x, 8x	NCBI	FM211885
Digitaria sanguinalis	2x, 4x, 6x, 8x	NCBI	FM211890
Echinochloa stagnina 1	4x, 6x, 12x, 14x	BioProject PRJNA395007	ESTA_c28431_g1_i1
Echinochloa oryzicola	4x	http://ibi.zju.edu.cn/RiceWeedomes/Echinochloa/	Contig44_pilon.31
Echinochloa oryzicola	4x	http://ibi.zju.edu.cn/RiceWeedomes/Echinochloa/	Contig244_pilon.619

Species	Ploidy Level	Resource	GENE/ACCESSION
Echinochloa oryzicola 1	4x	http://ibi.zju.edu.cn/RiceWeedomes/Echinochloa/	Contig317_pilon.71
Echinochloa oryzicola 2	4x	http://ibi.zju.edu.cn/RiceWeedomes/Echinochloa/	Contig1309_pilon.46
Echinochloa crus-galli	2x, 4x, 6x, 8x, 10x	http://ibi.zju.edu.cn/RiceWeedomes/Echinochloa/	Assembly
Echinolaena inflexa	10x	NCBI	FM211842
Echinochloa colona 1	6x	BioProject PRJNA395007	GFJI_GFJI01363984.1
Eleusine coracana GGLZ 1	2x, 4x	TSA database	GGLZ_GGLZ01024906.1
Eleusine coracana GGPD 2	2x, 4x	TSA database	GGPD_GGPD01027472.1
Eleusine indica 1	2x, 4x	NCBI	FM211843
Eleusine intermedia GGMC 1	2x	TSA database	GGMC_GGMC01029406.1
Eleusine multiflora GGLR 1	2x	TSA database	GGLR_GGLR01017951.1
Eleusine tristachya GGMD 1	2x	TSA database	GGMD_GGMD01031066.1
Echinochloa haploclada	2x	http://ibi.zju.edu.cn/RiceWeedomes/Echinochloa/	chr1.4620.mRNA1
Echinochloa haploclada	2x	http://ibi.zju.edu.cn/RiceWeedomes/Echinochloa/	chr4.1894.mRNA1
Enteropogon prieurii 2	NA	NCBI	FM211844
Enteropogon prieurii 1	NA	NCBI	FM211845
Enteropogon prieurii	NA	NCBI	FM211891
Entolasia marginata	NA	BioProject PRJNA395007	DN25362_c0_g1_i1
Eragrostis curvula 1	2x, 4x, 5x, 6x, 7x, 8x	Ensembl Plants	EJB05_07288
Eragrostis minor 1	2x, 3x, 4x, 6x, 8x	NCBI	FM211846

Species	Ploidy Level	Resource	GENE/ACCESSION
Eragrostis minor 2	2x, 3x, 4x, 6x, 8x	NCBI	FM211847
Eragrostis tef 1	4x	Ensembl Plants	Et_s3379-0
Eragrostis tenuifolia 1	NA	NCBI	FM211881
Eragrostis nindensis1	NA	NCBI	En_0044490-RA
Eragrostis nindensis 2	NA	NCBI	En_0008721-RA
Eragrostis nindensis 4	NA	NCBI	En_0041354-RA
Eragrostis nindensis 3	NA	NCBI	En_0073565-RA
Eriochloa nana	4x	NCBI	FR872782
Eriochloa nana	4x	NCBI	FR872783
Flaveria pringlei	4x	BioProject PRJNA395007	AB050473.Flaveria.pringlei
Flaveria trinervia	4x	NCBI	AB050472.Flaveria.trinervia
Flaveria trinervia	4x	NCBI	AB050471.Flaveria.trinervia
Garnotia stricta var. longiseta	4x	NCBI	MK737797
Glycine max	2x, 4x, 6x, 8x	BioProject PRJNA395007	Glyma04g09510.1
Guadua sp. Hodkinson s.n.	NA	NCBI	FM211848
Guadua angustifolia	2x	http://www.genobank.org/bamboo	Gan02440712exon(s)1446-63621992
Guadua angustifolia	2x	http://www.genobank.org/bamboo	Gan02201712exon(s)1825-73131986bp
Holcus lanatus	2x	NCBI	FM211849
Homopholis proluta	NA	BioProject PRJNA395007	HPRO_c15119_g3_i1
Hordeum vulgare	2x, 4x, 10x	Ensembl Plants	HORVU2Hr1G029160_GENOMIC

Species	Ploidy Level	Resource	GENE/ACCESSION
Hordeum vulgare	2x, 4x, 10x	Ensembl Plants	HORVU4Hr1G062440_GENOME
Hordeum vulgare GoldenPromise	NA	Ensembl Plants	HORVU.MOREX.r2.4HG0325280.1
Hordeum vulgare subsp. vulgare	NA	NCBI	AK362286
Hymenachne amplexicaulis	2x	BioProject PRJNA395007	HAMP_c10248_g1_i1
Hyparrhenia hirta	2x	NCBI	FM211818
Ichnanthus vicinus	2x, 4x	NCBI	FM211850
Imperata cylindrica	2x,4x,6x	NCBI	FM211882
Isachne mauritiana	2x	NCBI	FM211851
Jansenella griffithiana	2x, 4x	NCBI	MK737798
Lasiacis sorghoidea	2x, 4x	NCBI	FR872785
Lasiacis sorghoidea	2x, 4x	NCBI	FR872786
Lasiacis sorghoidea	2x, 4x	BioProject PRJNA395007	LSOR_c27006_g1_i1
Leersia perrieri	2x	Ensembl Plants	LPERR03G09790
Leersia perrieri	2x	Ensembl Plants	LPERR10G03780
Lepturus repens 1	2x	NCBI	FM211852
Megathyrsus maximus	NA	NCBI	AF532733
Megathyrsus maximus	NA	NCBI	FM211879
Megathyrsus maximus	NA	NCBI	FM211880
Megathyrsus maximus	NA	NCBI	FM211893
Megathyrsus maximus1	NA	BioProject PRJNA395007	GFVJ_GFVJ01102406.1
Melica uniflora	2x, 6x	NCBI	FM211853

Species	Ploidy Level	Resource	GENE/ACCESSION
Melinis minutiflora 1	2x	NCBI	FM211856
Melinis repens 1	2x	NCBI	FM211854
Melinis repens 2	2x	NCBI	FM211855
Merxmuellera macowanii	2x, 8x	NCBI	FM211869
Microlaena stipoides	2x	NCBI	FM211858
Miscanthus sacchariflorus	2x, 6x	NCBI	GCA_002993905
Miscanthus sinensis	2x, 4x	Phytozome	Misin02G396200/Misin01G412200
Musca acuminata	NA	Phytozome	GSMUA_Achr4G22070_001
Musca acuminata	NA	Phytozome	GSMUA_Achr8G18810_001
Olyra latifolia	2x	http://www.genobank.org/bamboo	scaffold63:372696:377542
Oplismenus hirtellus	6x, 8x, 10x	NCBI	FM211859
Oropetium thomaeum 1	2x	Phytozome	20150105_00740A
Orthoclada laxa	2x	NCBI	FM211860
Oryza barthii	2x, 3x	Ensembl Plants	OBART03G11030
Oryza barthii	2x, 3x	Ensembl Plants	OBART10G05150
Oryza brachyantha	2x	Ensembl Plants	OB03G21230
Oryza brachyantha	2x	Ensembl Plants	OB10G14040
Oryza brachyantha	2x	NCBI	XM_006649702
Oryza glaberrima	2x	Ensembl Plants	ORGLA03G0106000
Oryza glaberrima	2x	Ensembl Plants	ORGLA10G0042300
Oryza glumipatula	2x	Ensembl Plants	OGLUM03G10960

Species	Ploidy Level	Resource	GENE/ACCESSION
Oryza glumipatula	2x	Ensembl Plants	OGLUM10G05040
Oryza indica	NA	Ensembl Plants	BGIOSGA032182
Oryza indica	NA	Ensembl Plants	BGIOSGA032183
Oryza longistaminata	2x, 4x	Ensembl Plants	KN538688
Oryza longistaminata	2x, 4x	Ensembl Plants	KN540949
Oryza meridionalis	2x	Ensembl Plants	OMERI03G10310
Oryza nivara	2x	Ensembl Plants	ONIVA03G11670
Oryza nivara	2x	Ensembl Plants	ONIVA10G04640
Oryza punctata	2x, 4x	Ensembl Plants	OPUNC03G10660
Oryza punctata	2x, 4x	Ensembl Plants	OPUNC10G04650
Oryza rufipogon	2x, 4x	Ensembl Plants	ORUFI03G11350
Oryza rufipogon	2x, 4x	Ensembl Plants	ORUFI10G05420
Oryza sativa	2x, 3x, 4x	Phytozome	Os03g15050.1
Oryza sativa	2x, 3x, 4x	Ensembl Plants	Os03g0255500
Oryza sativa	2x, 3x, 4x	Ensembl Plants	Os10g0204300
Oryza sativa Indica Group	NA	NCBI	CT830933
Oryza sativa Japonica	NA	NCBI	LOC_Os03g15050.1
Oryza sativa Japonica Group	NA	NCBI	AF251066
Oryza sativa Japonica Group	NA	NCBI	AK102392
Oryza sativa Japonica Group	NA	NCBI	AK103839
Oryza sativa Japonica Group	NA	NCBI	XM_015758799

Species	Ploidy Level	Resource	GENE/ACCESSION
Oryza sativa Japonica Group	NA	NCBI	XM_015775203
Panicum hallii	2x, 4x	Phytozome	H02434.1
Panicum hallii	2x, 4x	Ensembl Plants	GQ55_9G538200
Panicum hallii	2x, 4x	Phytozome	PAHAL_9G527500
Panicum hallii	2x, 4x	NCBI	XM_025938884
Panicum laetum	NA	NCBI	FM211862
Panicum miliaceum	2x, 4x, 6x, 8x	NCBI	FM211863
Panicum parvifolium	2x, 4x	NCBI	FR872789
Panicum queenslandicum	NA	BioProject PRJNA395007	PQUE_c19232_g1_i1
Panicum virgatum	2x, 4x 8x, 10x	Phytozome	Pavir.Ia03881.1
Panicum virgatum	2x, 4x 8x, 10x	Phytozome	Pavir.Ib01078.1
Paspalum conjugatum	2x, 4x, 8x	NCBI	FM211866
Paspalum fimbriatum	2x, 4x	BioProject PRJNA395007	PFIM_c23689_g1_i1
Paspalum notatum	2x, 3x, 4x, 60x	TSA database	GFNR_GFNR01002556.1
Paspalum notatum	2x, 3x, 4x, 6x	TSA database	GFNR_GFNR01002558.1
Paspalum paniculatum	2x, 4x, 6x	NCBI	FM211884
Paspalum quadrifarium	2x, 3x, 4x, 6x	NCBI	FM211864
Paspalum quadrifarium	2x, 3x, 4x, 6x	NCBI	FM211865
Phragmites australis	2x, 8x, mixoploid	NCBI	FM211868
Phyllostachys edulis	4x	NCBI	FP097036

Species	Ploidy Level	Resource	GENE/ACCESSION
Poeae sp.	NA	BioProject PRJNA395007	PSSP_c29464_g1_i2
Populus trichocarpa	4x	BioProject PRJNA395007	POPTR_0007s14250.1
Populus trichocarpa	4x	BioProject PRJNA395007	POPTR_0002s10850.1
Puccinellia tenuiflora	2x, 4x	NCBI	GCA_012064385.1_evm.model.fragScaff_162.19
Raddia distichophylla	NA	China National Centre for Bio information (CNCB)	GCA_005191435.1_GWHTAAKD015025
Raddia guianensis	NA	http://www.genobank.org/bamboo#2	Rgu005627.1
Ricinus communis	2x	NCBI	XM_002528858.Ricinus.communis
Ricinus communis	2x	NCBI	XM_002509951.Ricinus.communis
Saccharum spontaneum	2x, 3x, 4x, 6x, 8x, 12x, ect.	Ensembl Plants	Sspon.01G0016840
Saccharum hybrid	NA	NCBI	SCSP803280_000008804.2
Sacciolepis indica	2x, 4x	NCBI	FM211870
Sacciolepis striata	2x, 4x	BioProject PRJNA395007	SSTR_c15512_g1_i1
Setaria barbata 1	4x, 5x	BioProject PRJNA395007	SBAR_c33190_g2_i1
Setaria italica 1	2x, 4x	Ensembl Plants	SETIT_034404mg
Setaria italica 3	2x, 4x	NCBI	XM_004984910
Setaria palmifolia 1	4x,6x	NCBI	FR845851
Setaria plicata 1	2x, 4x, 6x, 8x	NCBI	FM211861
Setaria viridis 1	2x, 4x	Ensembl Plants	SEVIR_9G469000v2
Setaria viridis 2	2x, 4x	NCBI	FM211886
Setaria viridis 3	2x, 4x	NCBI	XM_034717075

Species	Ploidy Level	Resource	GENE/ACCESSION
Setaria italica 2	2x, 4x	Phytozome	Seita.9G465500.1
Setaria virdis	2x, 4x	Phytozome	Sevir.9G469000
Solanum lycopersicum	2x,3x,4x	NCBI	NM_001247150.Solanum.lycopersicum
Sorghum bicolor	2x, 4x, 5x	Phytozome	Sobic.001G432800.1
Sorghum bicolor	2x, 4x, 5x	Ensembl Plants	SORBI_3001G432800
Sorghum bicolor	2x, 4x, 5x	NCBI	XM_021454674
Sorghum bicolor	2x, 4x, 5x	NCBI	XM_021454677
Sorghum bicolor	2x, 4x, 5x	BioProject PRJNA395007	Sb01g040720.1
Spinifex littoreus 1	2x	NCBI	FM211873
Sporobolus africanus 2	2x-4x	NCBI	FM211887
Sporobolus africanus 1	2x-4x	NCBI	FM211888
Sporobolus africanus	2x-4x	NCBI	FM211892
Sporobolus anglicus	NA	NCBI	FM211871
Sporobolus anglicus	NA	NCBI	FM211872
Sporobolus festivus 1	2x	NCBI	FM211874
Sporobolus festivus 2	2x	NCBI	FM211875
Sporobolus maritimus	NA	NCBI	FM211889
Sporobolus maritimus	NA	NCBI	FM211894
Sporobolus maritimus	NA	NCBI	GU204987
Sporobolus schoenoides	NA	NCBI	FM211834
Sporobolus stapfianus 1	NA	BioProject PRJNA395007	GFJP_GFJP01003961.1

Species	Ploidy Level	Resource	GENE/ACCESSION
Steinchisma sp.	NA	BioProject PRJNA395007	OSPP_c1172_g1_i1
Stipagrostis pennata	NA	NCBI	FM211876
Tenaxia disticha	2x	NCBI	FM211857
Theobroma cacao	2x, 4x	BioProject PRJNA395007	Glyma01g02330.1
Theobroma cacao cultivar Matina	2x, 4x	BioProject PRJNA395007	Glyma09g33650.1
Thinopyrum elongatum	2x, 4x	China National Centre for Bio information (CNCB)	CDS
Thyridolepis mitchelliana	2x	NCBI	FR872784
Thysanolaena latifolia	2x, 4x	NCBI	FM211877
Tristachya leucothrix	2x	NCBI	FM211878
Triticum aestivum	6x	Ensembl Plants	TraesCS4A02G083900
Triticum aestivum	6x	Ensembl Plants	TraesCS4B02G220200
Triticum aestivum	6x	Ensembl Plants	TraesCS4D02G220500
Triticum aestivum	6x	NCBI	AK450680
Triticum aestivum Cadenza	6x	Ensembl Plants	TraesCAD_scaffold_100752_01G000100
Triticum aestivum Claire	6x	Ensembl Plants	TraesCLE_scaffold_110547_01G000100
Triticum aestivum Paragon	6x	Ensembl Plants	TraesPAR_scaffold_091464_01G000200
Triticum aestivum Paragon	6x	Ensembl Plants	TraesPAR_scaffold_098135_01G000100
Triticum aestivum Robigus	6x	Ensembl Plants	TraesROB_scaffold_063847_01G000100
Triticum aestivum Weebill	6x	Ensembl Plants	TraesWEE_scaffold_088765_01G000100
Triticum dicoccoides	4x	Ensembl Plants	TRIDC4AG011450
Triticum dicoccoides	4x	Ensembl Plants	TRIDC4BG038930

Species	Ploidy Level	Resource	GENE/ACCESSION
Triticum turgidum	4x	Ensembl Plants	TRITD4Av1G036710
Triticum turgidum	4x	Ensembl Plants	TRITD4Bv1G133560
Triticum urartu	2x	Ensembl Plants	TRIUR3_23243
Urochloa panicoides 1	4x, 8x	NCBI	AF136161
Urochloa panicoides 2	4x, 8x	NCBI	UP09241
Urochloa villosa	4x	NCBI	FM211823
Urochloa villosa	4x	NCBI	FM211824
Urochloa villosa	4x	NCBI	FM211825
Zea mays	2x, 4x, 5x, 8x	Ensembl Plants	Ensembl-18_GRMZM5G870932_T01
Zea mays	4x, 8x	Ensembl Plants	Ensembl_GRMZM2G001696_T02
Zea mays	4x, 8x	Phytozome	PH207_Zm00008a000975
Zea mays	4x, 8x	Phytozome	PH207_Zm00008a035665
Zea mays	4x, 8x	Ensembl Plants	Zm00001d028471
Zea mays	4x, 8x	Ensembl Plants	Zm00001d047893
Zea mays	4x, 8x	NCBI	AB018744
Zea mays	4x, 8x	NCBI	AY109361
Zea mays	4x, 8x	NCBI	BT062880
Zea mays	4x, 8x	NCBI	BT062988
Zea mays	4x, 8x	NCBI	NM_001152706
Zea mays	4x, 8x	NCBI	NM_001309908
Zea mays	4x, 8x	NCBI	NM_001348550

Species	Ploidy Level	Resource	GENE/ACCESSION
Zea mays	4x, 8x	NCBI	NM_001348551
Zoysia japonica	4x	NCBI	AB199899
Zoysia japonica Nagirizaki 2	4x	http://zoysia.kazusa.or.jp/	sc00002.1.g07580
Zoysia japonica Nagirizaki 1	4x	http://zoysia.kazusa.or.jp/	sc00002.1.g07640
Zoysia matrella Wakaba 2	2x, 4x	http://zoysia.kazusa.or.jp/	sc02863.1.g00030
Zoysia matrella Wakaba 1	2x, 4x	http://zoysia.kazusa.or.jp/	sc02863.1.g00080
Zoysia matrella Wakaba 3	2x, 4x	http://zoysia.kazusa.or.jp/	sc09157.1.g00019
Zoysia matrella Wakaba 4	2x, 4x	http://zoysia.kazusa.or.jp/	sc09157.1.g00020
Zoysia pacifica Zanpa 2	4x	http://zoysia.kazusa.or.jp/	sc00001.1.g00260
Zoysia pacifica Zanpa 1	4x	http://zoysia.kazusa.or.jp/	sc00001.1.g00310

Chapter 4

4 Sequence similarity analyses suggest unidirectional lateral gene transfer from the grass *Themeda triandra* to the grass *Alloteropsis semialata*

Samuel G.S. Hibdige[1], Luke T. Dunning[1], Pascal-Antoine Christin[1]

Keywords: adaptation, evolution, genomics, horizontal gene transfer, lateral gene transfer, phylogenomics, Poaceae.

Affiliations: [1] Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom

Personal contributions: I designed the study with the help of my supervisors, performed all the analyses, and wrote the manuscript with the help of my supervisors.

4.1 Abstract

Lateral gene transfer (LGT) represents the movement of genetic material across species barriers by means other than sexual reproduction. While originally described in prokaryotes, LGTs have now been reported in a wide range of eukaryotes, and in plants seem especially abundant among grasses. The dynamics governing these gene movements remain, however, poorly studied. In this study we compare the genomes of two grass species, one of which was previously shown to have received LGT from the other, in order to test the hypothesis that LGT happens bidirectionally. We generate a *de novo* reference genome for the first species and compare it to an existing genome for the other species. Using similarity analyses, we detect 63 DNA segments that are at least 90% identical between these two species over more than 1,000 bp, with a similarity higher than expected based on other species from the same taxonomic groups. These segments include those previously identified using phylogenetic analyses, but also identify others spread across the genome. Most LGT seem to have moved in one direction, suggesting that LGT is largely unidirectional, with one species preferentially acting as the donor and the other as the recipient.

4.2 Introduction

Exchanges of genetic material among distinctly related species by means other than sexual reproduction with recombination, known as lateral gene transfer (LGT) or horizontal gene transfer (HGT), are well documented in prokaryotes. Such LGT have more recently been reported in a variety of eukaryotes (Keeling and Palmer 2008; Reynolds *et al*., 2018; Dunning *et al*., 2019; Xia *et al*., 2021), including diverse lineages of plants (El Bairouri *et al*., 2014; Li *et al*., 2014, 2018; Dunning *et al*., 2019; Wang *et al*., 2020). Among non-parasitic plants, LGT seems especially common in grasses (Mahelka *et al*., 2017, 2021; Dunning *et al*., 2019; Hibdige *et al*., 2021; Wu *et al*., 2022). It is now established that large genomic fragments can pass between species in a non-sexual manner, especially within the grass family. These genomic fragments transfers can include genes, some of which remain functional in the recipient species, with potential adaptive consequences (Olofsson *et al*., 2019; Phansopa *et al*., 2020). The amount and frequency of LGT seems to vary among groups, and comparative analyses have suggested that LGT was more frequent among closely related lineages. Furthermore rhizomatous species were more often the recipient of such exchanges than expected by chance (Hibdige *et al*., 2021). In the absence of a clearly identified transfer mechanism, the dynamics of these exchanges remain poorly understood. In particular, it is not known whether such exchanges happen in both directions, or whether some species are more likely to act as donor than as a recipient or vice versa.

It has been previously shown that one Australian accession of the grass *Alloteropsis semialata* has received a total of 59 protein-coding genes from various grass lineages (Dunning *et al*., 2019). This discovery was based on phylogenetic analyses of protein-coding genes, which can be reliably compared among grass species capturing the diversity of the family. The downside of this approach is that LGT discoveries are then restricted to protein-coding genes. By investigating the regions flanking these protein-coding LGT, Dunning *et al*., were able to show that non-coding DNA was included in the transferred fragments (Dunning *et al*., 2019). This conclusion was later confirmed for other donors and recipients (Hibdige *et al*., 2021). However, such secondary analyses cannot detect non-coding DNA transferred independently of the protein-coding genes. In addition, the fast turnover of non-coding regions in grass genomes means that such analyses can be applied only to recent LGTs where genome data for a close relative of both the donor and the recipient is known (Dunning *et al*., 2019; Hibdige *et al*., 2021). Testing for a reciprocity of LGT among donors and recipients therefore requires analyses of genomes for two taxa known to be involved in such transfers.

Most of the donors of LGT to *A. semialata* could not be identified to a species with confidence, because genome data is only available for a small fraction of the >12,000 grass species (Dunning *et al*., 2019). The most notable exception to this is the grass *Themeda triandra*, as LGT found solely in *A. semialata* from Australia are nested in phylogenetic trees within *T. triandra* individuals from Australia (Dunning *et al*., 2019). A total of eight protein-coding genes originating from *T. triandra* were detected in the genome of *A. semialata*. These genes belonged to two large fragments including large chunks of non-coding DNA. These analyses were based on low-coverage sequence datasets for *T. triandra,* and the absence of an assembled genome for this species prevented testing for potential DNA transfers in the opposite direction, from *A. semialata* to *T. triandra.*

In this paper, we test for the reciprocity of LGT by focusing on the pair of grasses composed of *A. semialata* and *T. triandra*. We generate a new nuclear genome assembly for *T. triandra* and then use a similarity-based approach to detect potential LGT. Our aims are (i) to validate the similarity-based approach by testing whether we can re-detect known LGT in *A. semialata*, (ii) to establish whether non-coding LGT is present in other parts of the genome of *A. semialata,* and (iii) to test whether DNA was also transferred from *A. semialata* to *T. triandra.*

4.3 Methods

4.3.1 Sampling strategy

We compared the genomes of *Themeda triandra* and *Alloteropsis semialata* to identify regions with unexpectedly high sequence similarity, which might be interpreted as resulting from lateral gene transfer (LGT). The chromosome-level genome of an Australian accession of *A. semialata* (accession AUS1) was generated previously (Dunning *et al*., 2019) and used here. Because high sequence similarity can be observed in ultraconserved genes (Reneker *et al*., 2012), we also estimated the similarity between *T. triandra* sequences and two other species from the tribe that contains *A. semialata* (Paniceae); *Panicum virgatum* and *Setaria italica* (Table 4.1, Figure 4.1). We then focus on segments largely more similar between *T. triandra* and *A. semialata* than between *T. triandra* and any of the other two Paniceae. Finally, to differentiate genes passed from *T. triandra* to *A. semialata* to those potentially passed from *A. semialata* to *T. triandra,* we estimated the similarity of the *T. triandra* segment to *Sorghum bicolor,* a species from the same tribe (Andropogoneae). A DNA segment that originated in *T. triandra* would be more similar to *S. bicolor* than to the Paniceae *S. italica* and *P. virgatum*, and conversely. We used published reference genomes (Table 4.1), except for *T. triandra*, which was sequenced and assembled here in house.

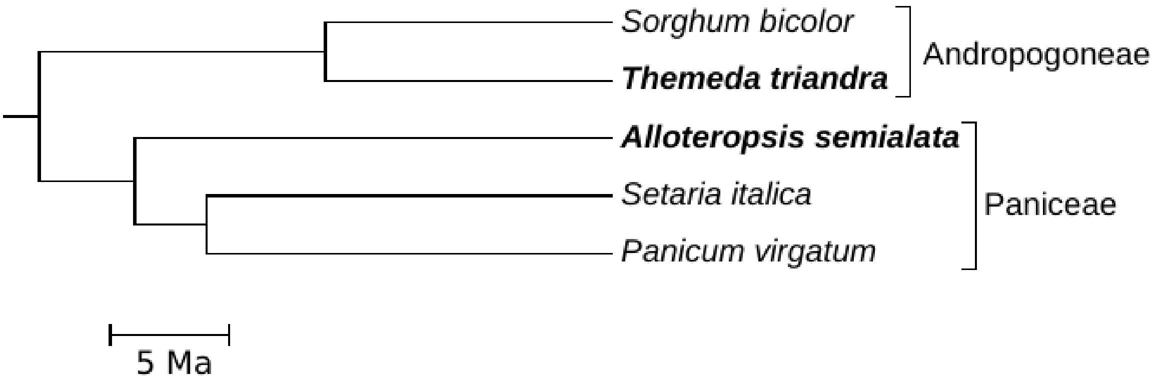


Figure 4.1: Relationships among the studied genomes. A time-calibrated phylogeny is shown for the five species analysed in this study. Relationships and divergence times (in million years; Ma) are based on Christin *et al*., (2012).

Table 4.1. Genome availability

Species	ID	Available from
Alloteropsis semialata	GCA_004135705.1	https://www.ncbi.nlm.nih.gov/
Panicum virgatum	GCF_016808335.1	https://www.ncbi.nlm.nih.gov/
Setaria italica	GCF_000263155.2	https://www.ncbi.nlm.nih.gov/
Sorghum bicolor	GCF_000003195.3	https://www.ncbi.nlm.nih.gov/

4.3.2 *Themeda triandra* reference genome assembly

A *de novo* reference genome assembly was generated for a *Themeda triandra* accession (TtPh16-4) collected in 2016 from the Carranglan region of the Philippines (15°56'35.8“ N 121°00'26.2“ E). A PacBio library was prepared by The University of Sheffield Molecular Ecology Laboratory, and sequenced on two PacBio Sequel SMRT cells. The PacBio data was cleaned and assembled using Canu v.2.0 (Koren *et al.*, 2017) with default parameters. Organelle genomes were then generated for the sequenced accession. The plastid genome was assembled using a genome walking implemented in Novoplasty (Dierckxsens *et al.*, 2016). The mitochondrial genome was manually assembled from the PacBio contigs. In brief, the complete set of mitochondrial genes was extracted from a *Sorghum bicolor* mitochondrial assembly (NC_008360.1) and used as a Blastn v.2.8.1 query to identify the top-hit TtPh16-4 contig for each gene. These contigs were then truncated to the matching regions, retaining the intergenic regions if multiple loci were present on a single contig. Finally, duplicated regions were removed and the remaining contigs concatenated into a single pseudomolecule with gaps represented by 100 Ns. The completeness of the TtPh16-4 mitochondrial genome was estimated using the MITOFY v.1.3.1 webserver (Alverson *et al.*, 2010).

The TtPh16-4 organelle genomes were used to mask organellar DNA in the Canu genome assembly prior to additional homology-based scaffolding. Contigs containing organellar DNA were first identified using Blastn, with a minimum alignment length of 1,000 bp and sequence similarity ≥ 99%. These scaffolds were then masked using RepeatMasker v.4.0.6 (Smit *et al.*, 2013) with the organelle sequences as a custom database. The organelle masked contigs were then scaffolded in relation to the genome of *Sorghum bicolor* (GenBank accession: GCA_000003195.3; McCormick *et al.*, 2018), a closely related grass from the same tribe (Andropogoneae), using RagTag v.2.1.0 (Alonge *et al.*, 2021). The TtPh16-4 genome assembly completeness was estimated using BUSCO v.3.1.0 (Simão *et al.*, 2015) with the poales_odb10 database, and by comparing the assembly size to the 1C genome size estimated for another individual collected from the same area (TtPh16-2) that was estimated by flow cytometry using the one-step protocol (Doležel *et al.*, 2007) with minor modifications (Clark *et al.*, 2016).

4.3.3 Identification of highly similar sequences

The genome assembly of *T. triandra* was cut into non-overlapping segments of 1,000, 5,000, 10,000 and 20,000 bp to identify the best size for LGT detection. This was limited by the N50 of the genome assembly of 22.45 kb. The 20,000 bp segment provided the best resolution and the rest were discarded. Segments with 10% or more ambiguous bases were discarded, and the others were used to assess the similarity to each of the four other genomes. Each 20,000 bp was successively compared to each of four genomes using Blastn 2.2.31+. In each case, the top match of each *T. triandra* was retained. Those *T. triandra* with a top match in *A. semialata* with a match length greater than 1,000bp and a pairwise identity greater than 90% were retained, and their pairwise identity to each of the three other references was recorded. 90% was chosen as a threshold to allow comparison to the mapping approach used in Dunning et al., 2019 as reads will map only if they are more than 90% similar to the reference (based on the default bowtie2 parameters and minimum scores for a valid alignment). We considered LGT candidates as those of the *T. triandra* segments with a pairwise identity to *A. semialata* that was at least 5% greater than their pairwise identity to both *P. virgatum* and *S. italica*. We then assigned LGT candidates to likely *A. semialata* or *T. triandra* origins based on their pairwise similarities to *S. bicolor* and the two Paniceae. Specifically, a segment was assumed to originate in *T. triandra* if it was more similar to *S. bicolor* than to both *P. virgatum* and *S. italica* by at least 1% identity. Conversely, segments more similar to both *P. virgatum* and *S. italica* than to *S. bicolor* by at least 1% identity were considered as originating in *A. semialata*. No likely origin was assigned to the other segments. The position of the LGT candidates along the chromosomes of *A. semialata* was plotted using R version 3.2.3.

4.4 Results

4.4.1 Reference genome statistics

We generated 20.93 Gb of PacBio subread data for the TtPh16-4 accession with an N50 read length of 5.61 kb. The initial Canu assembly consisted of 61,884 contigs with an N50 of 13.44 kb for a total length of 0.70 Gb, which is slightly below the genome size of Filipino *T. triandra* estimated with flow cytometry (0.84 Gb). We masked 3.08 Mb of organellar DNA before the final homology-based scaffolding in relation to the *S. bicolor* genome. In total, 19,639 contigs were scaffolded into 10 pseudo-chromosomes which had a combined length of 288.99 Mb (range 21.08 - 46.11 Mb). The Final genome assembly was composed of the 10 pseudo-chromosomes, the unplaced contigs and the organelle genomes. In total, there were 42,255 sequences, the N50 was 22.45 kb and the assembly size was 0.71 Gb (84.52% of the 0.84 Gb 1C flow-cytometry estimate genome size). The BUSCO poales_odb10 database contains 4,986 genes, of which 81.5% were complete (14% duplication). 2.4% were fragmented and 16.1% were missing in the TtPh16-4 reference genome.

4.4.2 Similarity analyses identify LGT

We considered 17,672 *T. triandra* segments of 20,000 bp, for a total length of 353,440 kb, representing about half of the assembled genome. Indeed, about half of the assembly length is included in contigs that are shorter than 20,000 bp of these 17,672 segments, 1,142 matched *A. semialata* on more than 1,000 bp with a pairwise identity above 90%, for a total match length of 3,177,687 bp. These numbers were reduced to 591 *T. triandra* segments and a total match length of 1,835,540 bp after considering only matches with a pairwise identity to *A. semialata* at least 5% greater, to avoid false positives or conserved regions, than to the two other Paniceae *P. virgatum* and *S. italica*. Multiple *T. triandra* segments can match the same position in *A. semialata*, and these were removed by considering only the *T. triandra* segment with the longest match in *A. semialata* among those with an overlapping match. The 591 segments were reduced to 63 unique matches, as most of the 591 segments represented different parts of the *T. triandra* assembly matching to the same part of the *A. semialata* genome as these likely represented repeats, transposable elements. These 63 unique matches represent a total of 190,520 bp. The longest match was 8,575 bp long, with a pairwise identity of 99.43% (highest observed pairwise identity). This segment corresponds to a gene encoding the enzyme phosphoenolpyruvate carboxylase previously shown to have been passed from *T. triandra* to *A. semialata* (Christin *et al*., 2012a; Dunning *et al*., 2019).

The position of the high-similarity segments was plotted along the nine *A. semialata* chromosomes (Figure 4.2). LGT candidates were detected on all nine chromosomes, but two clear clusters were apparent on chromosomes 4 and 7, respectively. These clusters correspond to the large multigene fragments previously identified as having been transferred from *T. triandra* to *A. semialata* (Dunning *et al*., 2019). The detection of many candidates in these regions validates our approach. In addition to these two fragments, seven other fragments have been passed from other Andropogoneae to *A. semialata* (Dunning *et al*., 2019), including the largest such fragments on chromosome 9 (Dunning *et al*., 2019). Of the LGT candidates detected here, only one corresponds to one of these fragments (on chromosome 9; Figure 4.2) and matches a coding sequence. These patterns confirm that similarity-based approaches are only useful to detect LGT when the donor is sampled, with only coding sequences maintaining sufficient similarity to compare among relatives of the direct donor.

Overall, our similarity analyses indicate that most of the DNA transferred between *T. triandra* and *A. semialata* is included in the two large fragments containing multiple protein-coding genes previously detected, but also reveal that shorter segments of laterally DNA are spread across the genome (Figure 4.2). These shorter segments likely consist of non-coding DNA, and in many cases of repeated sequences.

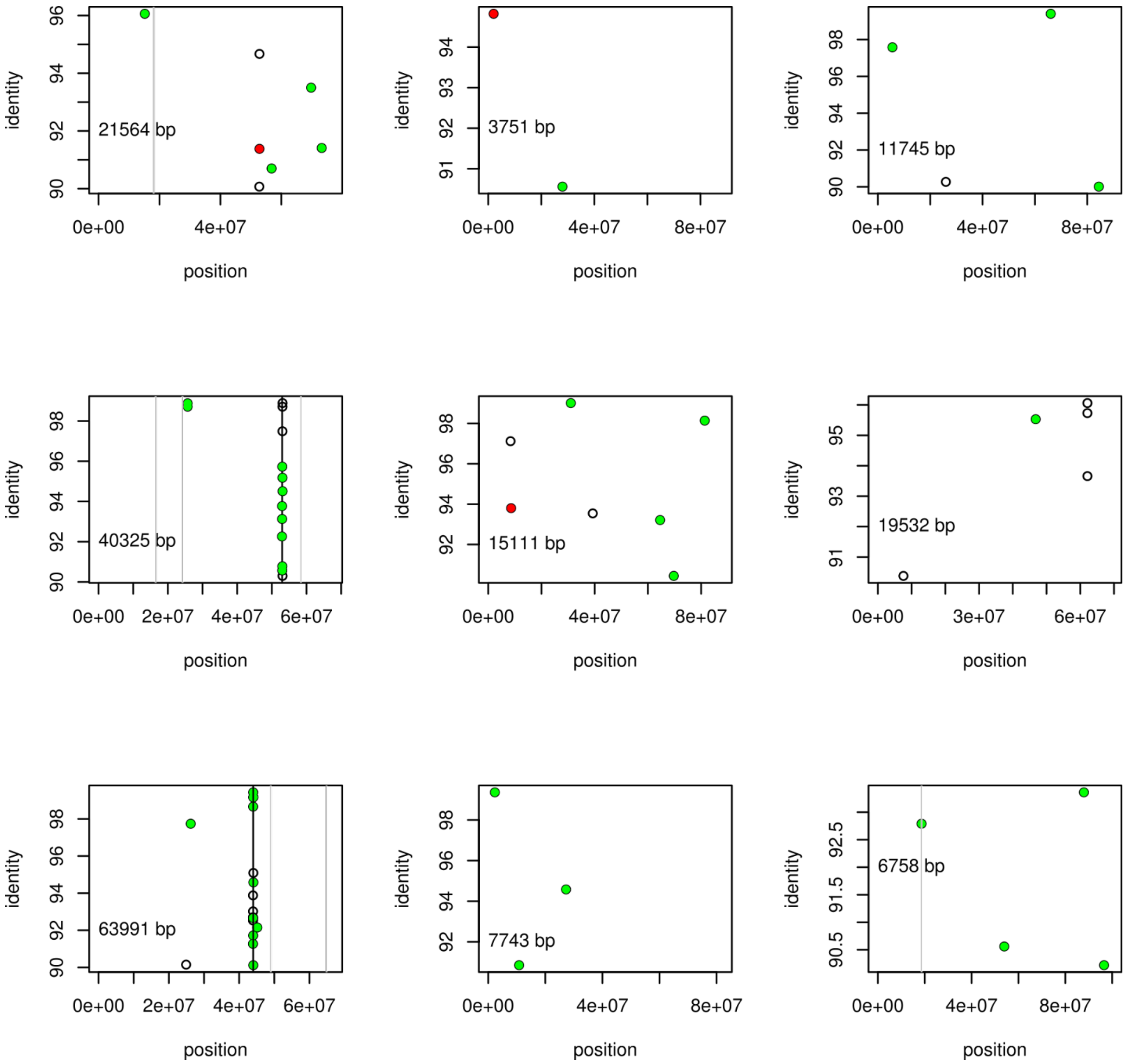


Figure 4.2: Distribution of LGT candidates along the genome of *Alloteropsis semialata*. The position along the nine chromosomes of *A. semialata* of each non-overlapping DNA segment presenting more than 90% identity between *A. semialata* and *Themeda triandra* on at least 90% is indicated, if the similarity between these two species was also greater by at least 5% than between *T. triandra* and other Paniceae. Segments with similarity patterns suggesting a transfer from *T. triandra* to *A. semialata* are in green, while those with similarity patterns suggesting a transfer from *A. semialata* to *T. triandra* are in red. Points in white are those for which the likely direction of the transfer could not be established. The total amount of DNA represented by the matches is indicated for each chromosome. The two black vertical bars indicate the position of the two previously detected fragments transferred from *T. triandra* into *A. semialata*. Vertical grey bars show the positions of LGT fragments received by *A. semialata* from other Andropogoneae grasses.

4.4.3 Genetic exchanges were mainly unidirectional

Of the 63 segments, 41 were more similar by at least 1% to *S. bicolor* than to both Paniceae, suggesting they were transferred from *T. triandra* to *A. semialata*. 1% was chosen as a threshold as these are comparisons to non-donor species at an already high percentage similarity. This includes most of the segments creating the two clusters on chromosomes 4 and 7 (Figure 4.2). For 19 segments, the pairwise identities were similar with *S. bicolor* and the Paniceae or *S. bicolor* was more similar than one of the Paniceae, but less similar than the other. Only three segments were more similar by at least 1% to both Paniceae than to Sorghum, as expected for a transfer from *A. semialata* to *T. triandra*. For two of these, homology among genomes was limited, preventing accurate alignment and phylogenetic analyses. The third segment could be compared among the different genomes on a region of about 2 kb. While it was indeed very similar to *A. semialata* and more similar to the Paniceae than to *S. bicolor*, it was even more similar to *Zea mays*, another Andropogoneae not included in the initial scan (Figure 4.3). This segment should thus be considered as likely originated in a relative of *T. triandra*, with the orthologous segment potentially lost in *S. bicolor* leading to a similarity estimate based on paralogs. Overall, these results strongly suggest that most, if not all, of the genetic exchanges went from *T. triandra* to *A. semialata*. We conclude that the lateral gene transfers between these two species were mostly unidirectional.

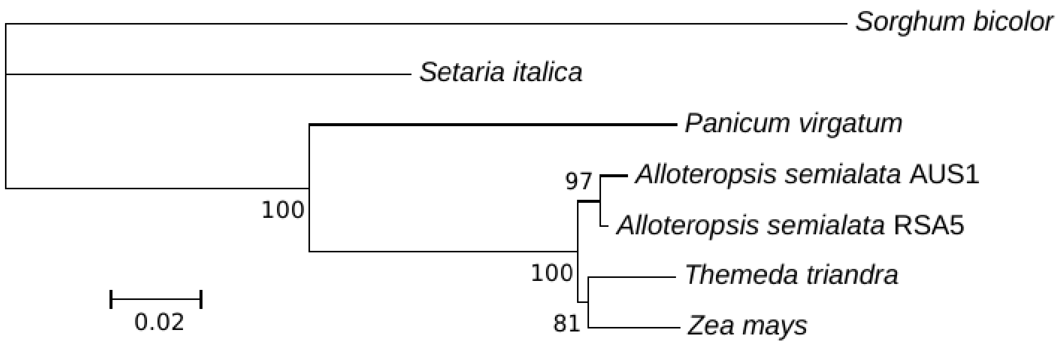


Figure 4.3: Phylogenetic tree inferred for one of the LGT candidates. The *Themeda triandra* sequence potentially acquired from *A. semialata* on chromosome 5 (Figure 4.2) and the matching *A. semialata* sequence were extracted. Homologous fragments were then retrieved from the genomes of *Panicum virgatum*, *Setaria italica,* and *Sorghum bicolor* (Table 4.1), in addition to *Zea mays* (another Andropogoneae; genome RefGen v4 from https://phytozome-next.jgi.doe.gov/) and another *A. semialata* accession (accession RSA5; from Raimondeau *et al*., unpublished). The genomic fragments were aligned with Muscle (Edgar 2004). The alignment was truncated to a 1914bp segment presenting good homology among the different species, and a phylogenetic tree was inferred with PhyML (Guindon *et al*., 2010), a HKY+G substitution model, and 100 bootstrap pseudoreplicates. Bootstrap support values are indicated near nodes.

4.5 Discussion

4.5.1 Similarity analyses can identify LGT among distinct species

The detection of lateral gene transfers (LGT) always rely on unexpected similarity between a gene found in one species and homologs found in a distant species. When phylogenetic trees can be inferred, they can be used to confirm that the observed similarity is due to a gene history that differs from the species tree and not other processes leading to high conservation of sequences over large evolutionary scales (Dunning *et al*., 2019). However, phylogenetic trees cannot always be inferred, and many studies have relied on direct comparisons of pairwise similarities to detect LGT (Cai *et al*., 2021; Park *et al*., 2021a, 2021b). Such analyses however require a null model to which observed similarities can be compared, and in this study, we opted to compare the observed similarity between the potential recipient and donors to those observed with other members of the groups containing either the recipient or the donor. The weakness of such analysis is that the results depend on arbitrary thresholds (Park *et al*., 2021a). Nevertheless, we are able to re-identify here LGT regions that were previously revealed using phylogenetic analyses of protein-coding genes (Figure 4.2). Importantly, LGT regions received from other species from the same group were not generally re-identified here (Figure 4.2), confirming that similarity-based approaches work only when genome data are available for very close relatives of the actual donor.

Similarity based analysis do confer advantages over a phylogenetic approach as large portions of the genome cannot be properly evaluated. First, in old groups such as grasses non-coding regions undergo rapid turnovers and cannot be reliably aligned. Only coding sequences can be analysed in such a manner allowing comparisons to only closer relatives. In a similarity approach as long as you have the donor, even regions of rapid turnover can be compared without the knowledge of coding regions. As a result even unannotated genomes can be used. Second, phylogenetic analyses cannot be used to reliably infer correct relationships on a large proportion of protein coding genes. Indeed, large numbers of genes that were too short and those that are present in an insufficient number of species are excluded in such analyses (Hibdige et al., 2021). Provided the genome is well resolved with few ambiguous bases in the intergenic regions, smaller genes will be detected in similarity based approaches as the flanking regions can be acquired in the LGT (Dunning et al., 2019). If LGTs are small they may not be detected by a similarity based approach if the window they are found within splits the LGT, this may be resolved by using a sliding window approach to dividing the genome.

Alternative methods for such analysis might include iterative sampling of subtrees throughout the genome as used in topology weighting (Martin and van Belleghem, 2017). This could be used to sample the genome in distinct or sliding windows as used in our similarity based approach. The downsides of this are that the creation of a subtree for a window takes longer than the equivalent similarity analysis and will need further processing to detect discordant trees as used in Hibdige et al., 2021. This would then require further checks to ensure that the discordant trees were not due to factors other than LGT. As such analysis would include intergenic regions that contain repeats and transposable elements, the incidents of false positives and manual vetting would be higher and take significantly longer than when using phylogenetic approaches on purely coding regions (Hibdige et al., 2021).

Our analyses detected sequences with high similarity between *T. triandra* and *A. semialata* spread across the genome of *A. semialata* (Figure 4.2). Such DNA might have been received independently from the large genomic blocks containing protein-coding genes. Alternatively, they could have originated in these genomic fragments, but have been subsequently moved to other parts of the genome. Establishing the exact history of these DNA sequences requires comparative analyses of a large number of genomes for both the donor and the recipient, which might soon be available as studies of *T. triandra* and *A. semialata* continue expanding the sampling for these grasses (Dunning *et al*., under review; Raimondeau *et al*., unpublished).

4.5.2 Lateral gene transfers seem unidirectional among these two species

The grass *T. triandra* has given at least eight genes to the Australian *A. semialata,* as part of two large genomic blocks (Dunning *et al*., 2019). In addition, African *T. triandra* seem to have independently provided protein-coding genes to African *A. semialata* (Raimondeau *et al*., unpublished). These patterns suggest that the two species possess features that make genetic exchanges possible. However, while our analyses confirm transfer of DNA from *T. triandra* to *A. semialata,* only three DNA segments present patterns that are compatible with a reverse movement from *A. semialata* into *T. triandra* (Figure 4.2). The history of these three segments is difficult to confirm as non-coding DNA is generally hard to align among distant relatives, but at least one of them seems to actually result from a transfer from Andropogoneae into *A. semialata* (Figure 4.3). We therefore conclude that the genetic exchanges were mainly unidirectional.

The unidirectionality of the LGT suggests that while *T. triandra* seems prone to give genes and *A. semialata* seems prone to receive these genes, the reverse is not true. Biases in the direction of genetic exchanges could result from different factors. First, *A. semialata* might be more permeable to foreign genetic material, as suggested by the comparatively high number of LGT detected in this species (Hibdige *et al*., 2021). Second, the amount of pollen released might vary among the two species, both as a function of their relative abundance and of their pollen production per individual. If LGT occurred during illegitimate pollination, as hypothesized (Wickell and Li, 2020), the species producing more pollen would be more likely to act as LGT donor. Indeed, a species colonising a new area may be more likely to acquire LGT via illegitimate pollination due to a lower relative abundance of their pollen compared to established species with large populations. In addition, resident species may possess locally adapted genetic variation that will be more advantageous to the alien species colonising the novel habitat. This may bias the apparent direction of transfer as these are more likely to be maintained than maladapted variants passed in the other direction. As accession and cultivar data for individual species grow, it will be possible to identify LGT patterns across its current biogeographical ranges. Third, the probability of retaining LGT might vary as a function of the species demography. For instance, species with small population sizes might be more likely to fix neutral LGT under drift and therefore act as a LGT recipient. Precise estimates of population sizes and pollen production are lacking for *T. triandra* and *A. semialata,* but detailed field surveys in the future might help explain the bias observed here.

A potential cause of the unidirectionality of the LGT observed may be due to the inability of *T. triandra* to produce rhizomes. If LGT was facilitated by a natural rhizome-root graft, genetic material could be preserved in the vegetative rhizome tissue. The same would not occur in the instance of LGT into the root of *T. triandra,* as roots from this species are not a source of vegetative growth. DNA from *A. semialata* landing in the roots of *T. triandra* would therefore not be transferred to follow-up generations.

4.5.3 Conclusions

In this study, we generate a reference genome for the grass *Themeda triandra* and use a similarity-based analysis to identify potential DNA segments transferred between this species and the grass *Alloteropsis semialata*. Our analyses re-identify previously detected LGT between these two species, supporting the value of the approach. We further detect potential LGT in other parts of the genome. However, the vast majority of the LGT seem to have moved from *T. triandra* into *A. semialata*, with only few candidates for a reverse movement. Our investigations therefore suggest that LGT between two species can be unidirectional, potentially because of the population sizes and reproductive systems of the two species.

4.6 Acknowledgements

This work was funded by a Royal Society Research grant (RGF\EA\180247). P.-A.C. is supported by a Royal Society University Research Fellowship (URF\R\180022). L.T.D is supported by a Natural Environment Research Council Independent Research Fellowship (NE/T011025/1).

Chapter 5

5 General Discussion

5.1 Using phylogenetic trees to detect LGTs

Lateral gene transfer or horizontal gene transfer (LGT/HGT) in eukaryotes has been a controversial topic that has gained more interest in recent years (Martin, 2017). Although documented in bacteria since the early 20th century (Griffith, 1928; Freeman, 1951), LGT was thought to be unique to single cell organisms that can easily pick up genetic material through various known mechanisms and incorporate it into their germ line (Freeman, 1951; reviewed in von Wintersdorff *et al*., 2016). Because such transfer is theoretically more complex in multicellular organisms, LGT was thought to be rare if existing at all in this group.

Due to huge advancements in sequencing technology and price decreases, the quality and quantity of genetic resources has increased at a rapid rate in the last decade. As the number of published genomes increases, especially within some groups of organisms such as with grasses due to their economic importance, further in-depth comparisons between species and identification of unexpected patterns becomes possible. In past analyses based on fragmented and poorly resolved genome information, phylogenetic or similarity patterns compatible with LGT were difficult to interpret. Indeed, without due diligence, such patterns might be interpreted as artefactual, potentially resulting from contamination or paralogy issues. In the cases of prokaryotes to eukaryotes transfers, the possibility of DNA contamination is especially problematic, as bacteria are present within and around other organisms.

One particularly high profile example of DNA contamination leading to an erroneous conclusion of eukaryotic LGT was offered by the initial assembly and publication of the tardigrade genome. Its original assembly and analysis led the authors to conclude that 16% of the tardigrade genome originated from LGT of diverse origins (Boothby *et al*., 2015). After scrutiny from other research groups, it became apparent that the high levels of LGT were likely due to contamination and the true figure was much lower (Arakawa, 2016; Bemm, Weiß, Schultz and Förster, 2016, Koutsovoulos *et al*., 2016). This problem and the ensuing controversy led to questions about the claims of LGT in eukaryotes and the patterns we should be expecting to see if eukaryotic LGT was really occurring (Martin, 2017). The rapid accumulation of high-quality genomic data for various taxonomic groups has however led to many solid examples of eukaryotic LGT (Li *et al*., 2014; Vallenback *et al*., 2008; Christin *et al*., 2012b; Prentice *et al*., 2015; Mahelka *et al*., 2017, 2021; Dunning *et al*., 2019), and in many cases, artefacts such as contamination and paralogy problems can be ruled out. The detection of LGT remains however challenging, and methodological innovations are required, both to identify LGT candidates and to validate them.

Many scans for LGT, especially in prokaryotes, are based on similarity indexes, with high sequence similarity between genes belonging to distant lineages interpreted as evidence for LGT (Ma *et al.*, 2022). Other patterns can however create high similarity among distant genes, and while multiple species can be incorporated in the analyses to assess the expected similarity (see Chapter 4), the precise history of genes is better inferred with phylogenetic trees based on a dense species sampling (Chapter 2). Early discoveries of LGT among grasses were indeed made incidentally, during analyses of gene phylogenetic trees (e.g. Christin *et al*., 2012a). Focusing on *Alloteropsis semialata*, a species in which the incidental LGT were discovered, our research group developed a pipeline to first identify LGT candidates based on similarity analyses and then validate them with phylogenetic trees (Dunning *et al.*, 2019). Analyses of sequencing replicates were further used to rule out contamination, while analyses of long sequencing reads allowed confirmation that the foreign DNA was really integrated in the chromosomes of the recipient (Dunning *et al*., 2019). These initial analyses were however labour intensive, and the focus on a single species prevented assessing the frequency of LGT across the group.

I improved the LGT detection pipeline to detect any LGT existing in the genomes of 17 different grasses (Chapter 2). The main methodological innovation was the removal of similarity-based analyses, so that phylogenetic trees were inferred for all genes in the analysed genomes. This allowed us to identify a total of 170 genes across 13 of the 17 genomes available at the time (Chapter 2). Our approach was purposefully designed to minimise false positives and reduce the effects of processes such as hybridisation and incomplete lineage sorting, and therefore the results we see are likely to be a very conservative estimate of the LGT present in grasses. The method relies on the repeated sampling of grass clades and as such, the analyses can be repeated as more grass genomes become available. As it stands however we were unable to detect LGT from about 30% of grass clades due to the lack of genomic resources. We do however demonstrate that LGT is prevalent across the entire family, and not limited to any life strategy or phylogenetic group.

A major downside of using phylogenetic analyses to detect LGT is that most of the genome cannot be properly evaluated. First, only coding sequences can be reliably aligned across old groups such as grasses, as their other genomic partitions undergo rapid turnovers only allowing comparisons to closer relatives. Second, phylogenetic analyses cannot be used to reliably infer correct relationships on a large proportion of protein coding genes. Indeed, we excluded genes that are too short and those that are present in an insufficient number of species (Chapter 2). Third, insufficient species sampling numbers within grass clades prevented examination of LGT among members of the same clade, with the exception of within Paniceae. Our filters based on statistical support, further excluded all genes that were insufficiently informative. The comparative analyses presented in Chapter 2 still provide estimates that can be compared among species, provided a similar proportion of genes cannot be statistically evaluated in all of them. Regardless of our method only detecting protein-coding gene transfers, it is known that non-coding fragments can occasionally be moved across grasses (El Baidouri *et al*., 2014; Park *et al*., 2021). Indeed, our analyses of regions flanking LGT detected such regions in several species (Chapter 2), but unlinked DNA would be missed in such scans.

To detect any kind of LGT, potentially including non-coding DNA, we opted for another set of analyses based purely on pairwise similarity (Chapter 4). We did re-identify previously discovered protein-coding LGTs, but also detected other types of LGT candidates spread across the genome. By definition, reliable phylogenetic trees cannot be inferred for most of these, so other processes are difficult to rule out. In addition, similarity analyses outside of protein-coding genes can only be used for recent LGT and when genomes are available for close relatives of both the donor of the recipient (Chapters 2 and 4). The existing methods therefore allow either the identification of LGT across large evolutionary distances based on a fraction of protein-coding genes (Chapters 2 and 3) or the detection of all types of LGT, but on very restricted evolutionary scales (Chapter 4). The accumulation of genomic data for species covering the diversity of grasses would solve some of these limitations, and further methodological improvements combining phylogenetic and similarity might in the future infer all types of LGT over large evolutionary scales.

5.2 Some groups are most likely to undergo LGTs

Following the discovery of LGTs in some plants, the obvious question became whether such a phenomenon is widespread or concerns only some specific species. In the case of parasitic plants, LGTs have been discovered in multiple lineages (Xi *et al*., 2012; Vogel *et al*., 2018; Yang *et al*., 2019; Yoshida *et al*., 2019), suggesting that it is frequent in this particular lifestyle. In the absence of a known mechanism, whether all members of non-parasitic groups are subject to LGT remains unknown.

My scan of multiple grass genomes for LGT has provided the first answer to this question, whilst at the same time indicating some hints about potential mechanisms. While LGT were detected in 13 out of 17 analysed grasses, we showed that increased phylogenetic distance decreased the instances of LGT acquisition (Chapter 2). This indicates that either the mechanism of transfer is more readily facilitated by relatedness, or that once acquired, it is easier to co-opt genes that have more similar regulatory mechanisms. However, LGT was still present at detectable levels across the span of the grasses, suggesting the adaptive advantage of laterally acquiring genes can be worth the cost associated with accommodating new genes in the recipient genome.

We further showed that biogeography may play an important role in LGT, illustrated by the fact that two members of the same genus showed drastically different acquisition patterns. *Panicum virgatum* and *Panicum hallii* acquired the majority of their LGT's from Andropogoneae (81%) and Chloridoideae (79%), respectively. This pattern mirrors the dominant grassland type in which each of the sequenced individuals occurs (Lehmann *et al.,* 2019). The importance of biogeography is further illustrated by the case of *Dichanthelium oligosanthes,* the only grass in our data set that showed a transfer from Pooideae to Paniceae, two groups that diverged more than 50 million years ago. *Dichanthelium oligosanthes* is a frost tolerant grass that inhabits colder areas than other Paniceae where it likely co-occurs with Pooideae. Overall however, quantifying the effect of biogeography remains difficult as one not only needs accurate spatial maps of species ranges, but historical ones as well, especially in the case of older LGT. Nevertheless, when species coexist with other distantly related but dominant grasses, LGT still occurs albeit at lower levels. Many of the protein coding LGTs identified are poorly known, and their exact function generally remains unidentified. As our knowledge of the function of these genes grows, it will become possible to elucidate the adaptive advantage of LGT, especially when tied to the biogeography.

Further understanding of biogeography could help elucidate the patterns observed in Chapter 2. In instances where the same gene has been laterally acquired multiple times such as in Chapter 3, the environment where the LGT likely took place could be examined. Are we seeing LGT from more closely related species because they are more likely to grow together? If there are multiple species with the same advantageous gene within the same environment, has the LGT recipient received the gene from the species it is most closely related to, or the species that has the higher abundance? The more instances of the same gene being laterally acquired in different species, the more this can be examined.

Besides the importance of phylogenetic distance and biogeography, the comparative analyses of grass genomes suggested that rhizomatous species are more prone to LGT (Chapter 2). While this trend needs to be confirmed with a denser species sampling, it suggests that rhizomes and other structures sustaining vegetative growth favour the lateral acquisition of genes. One possibility is that these structures allow interspecific inosculation, which, if confirmed, could allow gene movements as demonstrated in the case of grafts (Stegemann and Bock, 2009; Stegemann *et al*., 2012). Interspecific root grafts can be observed in nature (Graham and Bormann, 1966) and while grafting was assumed to be impossible in grasses and other monocots, this assumption was recently refuted (Reeves *et al*., 2022). Because grasses are often in close interspecific associations, such root-to-root contacts could be frequent. Such a mechanism is unlikely to account for all LGT, as the phenomenon can also be observed in non-rhizomatous species (Chapter 2). It is likely that multiple mechanisms can be involved, some of which might occur in all grass species (e.g. illegitimate pollination), while others might be restricted to some functional types (e.g. inoculation). This would in fact mirror the multitude of HGT mechanisms seen within bacteria.

However, if rhizomes are a factor that increases incidents of LGT by acting as an additional interface, this would pose the question why would LGT be unidirectional as observed in Chapter 4. In Chapter 4, we capitalise on genomic resources we generated for both the donor (*Themeda triandra*) and the recipient (*Alloteropsis semialata*) of previously identified LGTs (Dunning *et al*., 2019). This exceptional resource allows us to ask whether species that are often identified as the recipient of LGT (Chapter 2) also act frequently as the donor of such gene transfers. Our analyses re-identify previously detected LGT between these two species, and further detect potential LGT in other parts of the genome (Chapter 4). However, the vast majority of the LGT seem to have moved from *T. triandra* into *A. semialata*, with only few candidates in a reverse movement. Our investigations therefore suggest that LGT between two species can be unidirectional, potentially because of the population sizes and reproductive systems of the two species. Perhaps this is a question of biogeography and species abundance, whereby a skewed population dominated by *Themeda triandra* statistically favours LGT in one direction. An alternative explanation is that if *Alloteropsis semialata* was colonising an area where *T. triandra* was present, then perhaps any genomic fragments would only be favourable to the coloniser as any LGT in the opposite direction would not necessarily be optimised for the conditions of the area. As accession and cultivar data for individual species grow, it will be possible to identify LGT patterns across its current biogeographical ranges.

The research presented in Chapter 3 therefore indicates that the species more likely to act as LGT recipients are not necessarily those more likely to act as LGT donors, supporting a complex dynamics of unidirectional gene exchanges among lineages of grasses. If the trend of unidirectional gene transfer is repeated across many pairs of donors and recipients, LGT may be seen as more akin to genetic parasitism by the recipient as there is no benefit to the donors if the transfers are one way.

5.3 Some genes are more likely to be involved in LGT

Besides the distribution of LGTs among lineages and species, the discovery of multiple instances of gene transfers leads to question whether some genetic elements are more likely to be exchanged than others. For example, previous investigations have suggested that genetic elements that are inherently prone to move within genomes (i.e. transposable elements) are also more often exchanged among species (El Baidouri *et al*., 2014; Park *et al*., 2021), although such conclusions might be affected by the detection method (see above). If true, such a pattern would point to mechanistic biases among genomic partitions. It is however equally possible that the likelihood of transfer, and more specifically of post-transfer retention, varies among genes as a function of their adaptive value. Indeed, a random genetic element landing in the genome of another species will be subjected to drift and potentially negative selection, and would therefore be unlikely to rise to fixation and be passed to future generations. Conversely, a genetic element providing an advantage to the recipient species would be subjected to positive selection, helping its spread within the recipient species (Olofsson *et al*., 2019). These dynamics have been discussed by analysing the history and functional impacts of genes within one recipient species (e.g. Olofsson *et al*., 2019; Phansopa *et al*., 2020), but whether some protein-coding genes are more prone to LGT remains speculative.

In Chapter 3, we reanalysed a gene previously shown to be involved in LGT among some grasses (Christin *et al*., 2012a; Dunning *et al*., 2019). Importantly, the phylogenetic tree for this gene encoding phosphoenolpyruvate carboxykinase (PCK) suggested that the gene had been independently passed to different grass lineages (Dunning *et al*., 2019). Using a larger sampling of grass genomes, we confirmed that *pck* genes have been independently transferred from some Cenchrinae species to both *A. semialata* and species from the *Echinochloa* genus belonging to the same tribe (Chapter 3). Reanalysis of a previously assumed deep duplication in the Chloridoideae subfamily of grasses (Christin *et al*., 2008) further showed a minimum of four additional lateral gene transfers, indicating that *pck* genes have in total been transferred from at least two groups of donors to at least six different recipients (Chapter 3). This exceptional case of repeated LGT likely results from the importance of *pck* genes for C_4 photosynthesis, which is used by all recipient species, and the necessity to duplicate *pck* genes before their recruitment into the C_4 pathway. Together, these two features mean that LGT provides, in this case, an alternative to gene duplication followed by neofunctionalization, conferring an advantage to the recipient of such gene transfers.

While my work suggests that some genes can be especially prone to LGT because of the advantage they confer to the recipient, future analyses will need to establish whether there are genome-wide patterns that explain the identity of genes successfully transferred among grass species. One possibility is that those genes that represent a functional novelty that arose after the split of the donor and recipient are more likely to be selectively retained following the transfers. The species identified as frequently involved in the LGT as either donor or recipients (Chapters 2 and 4) would constitute the perfect study system for comparative transcriptomics, asking whether genes that diverged in expression between the donor and recipient are more likely to be successfully retained following a lateral gene transfer. This could be paired with analyses of the fate of the LGT fragments in the recipient species, examining whether one advantageous gene causes an entire fragment to be maintained and whether the expression profiles of the other genes of the fragment are silenced across populations.

5.4 Crops are genetically modifying themselves

The presence of LGT within crop species is in itself interesting as this could have important implications to genetically modified (GM) crops. It has previously been shown that the sweet potato genome contained expressed *Agrobacterium tumefaciens* T-DNA inserts that were hypothesised to have been selected for during domestication (Kyndt *et al*., 2015). We confirm the presence of LGT in other crops species, in this case belong to the grass family (Chapter 2), reinforcing the idea that insertion of functional genes is no more unnatural than selective breeding. However, T-DNA inserts are limited in the size of genetic material that can be transferred. The fragments observed in grasses are a much larger size than *Agrobacterium* could account for, demonstrating there must be an alternative mechanism for grass-to-grass LGT.

The fact that genetic material can be transferred via means other than sexual reproduction does have implications for genetic escape into the environment. This means that if pesticide producing crops are engineered, it would be possible for these genes to be transferred to wild relatives by means other than hybridization and subsequent introgression. Unless a mechanism is identified, such escape would be hard to prevent. One possible solution to reduce genetic escape would be to grow grass crops in areas where only distantly related wild relatives exist, as LGT appears less frequent among distant lineages (Chapter 2). If the GM nature of the crop was related to stress tolerance, it would be preferable to use genes from the wild grasses where it is to be grown. Not only would this negate any effect of escape into wild grasses but it would also be easier to use genes from closely related species that are adapted to the stress existing within the environment.

In conclusion, LGT in grasses is both widespread and adaptive. How significantly it contributes as a driving force in evolution is yet to be fully established. However, I don’t think we have seen the tip of the iceberg yet.

References

Allender, C., 2011. The second report on the state of the world's plant genetic resources for food and agriculture. Rome: food and agriculture organization of the united nations. *Experimental Agriculture*, 47(3), pp.574-574.

Alonge, M., Lebeigle, L., Kirsche, M., Aganezov, S., Wang, X., Lippman, Z.B., Schatz, M.C. and Soyk, S., 2021. Automated assembly scaffolding elevates a new tomato system for high-throughput genome editing. *bioRxiv*. doi:10.1101/2021.11.18.469135.

Alvarez, C.E, Bovdilova, A., Höppner, A., Wolff, C.C., Saigo, M., Trajtenberg, F., Zhang, T., Buschiazzo, A., Nagel-Steger, L., Drincovich, M.F., Lercher, M.J. and Maurino, V.G., 2019. Molecular adaptations of NADP-malic enzyme for its function in C_4 photosynthesis in grasses. *Nature Plants*. 5(7), pp.755-765. doi:10.1038/s41477-019-0451-7.

Alverson, A.J., Wei, X., Rice, D.W., Stern, D.B., Barry, K. and Palmer, J.D., 2010. Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). *Molecular Biology and Evolution*, 27(6), pp.1436-1448. doi:10.1093/molbev/msq029.

Andersson, J.O., 2005. Lateral gene transfer in eukaryotes. *Cellular and Molecular Life Sciences Cell Molecular Life Sciences* 62(11), pp.1182-1197. doi:10.1007/s00018-005-4539-z.

Arakawa, K., 2016. No evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proceedings of the National Academy of Sciences of the United States of America*, 113(22), pp. 3057. doi:10.1073/pnas.1602711113.

Aubry, S., Brown, N.J. and Hibberd, J.M., 2011. The role of proteins in C_3 plants prior to their recruitment into the C_4 pathway. *Journal of Experimental Botany*, 62(9), pp.3049-3059. doi:10.1093/jxb/err012.

Barrett R.D.H. and Schluter D., 2008. Adaptation from standing genetic variation. *Trends in Ecology & Evolution* 23(1), pp.38-44. doi:10.1016/j.tree.2007.09.008.

Beiko, R.G., Harlow, T.J. and Ragan, M.A., 2005. Highways of gene sharing in prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(40), pp.14332-14337. doi:10.1073/pnas.0504068102.

Bellasio, C. and 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 Physiology*, 164(1), pp.466-480.doi:10.1104/pp.113.228221.

Bemm, F., Weiß, C.L., Schultz, J. and Förster, F., 2016. Genome of a tardigrade: Horizontal gene transfer or bacterial contamination?. *Proceedings of the National Academy of Sciences of the United States of America*, 113(22), E3054-E3056. doi:10.1073/pnas.1525116113.

Bennetzen, J.L., Schmutz, J., Wang, H., Percifield, R., Hawkins, J., Pontaroli, A.C., Estep, M., Feng, L., Vaughn, J.N., Grimwood, J., Jenkins, J., Barry, K., Lindquist, E., Hellsten, U., Deshpande, S., Wang, X., Wu, X., Mitros, T., Triplett, J., Yang, X., Ye, C.U., Mauro-Herrera, M., Wang, L., Li, P., Sharma, M., Sharma, R., Ronald, P.C., Panaud, O., Kellogg, E.A, Brutnell, T.P., Doust, A.N., Tuskan, G.A., Rokhsar, D. and Devos K.M., 2012. Reference genome sequence of the model plant *Setaria*. *Nature Biotechnology*, 30, pp.555-561. doi:10.1038/nbt.2196.

Bergthorsson, U., Richardson, A.O., Young, G.J, Goertzen, L.R. and Palmer, J.D., 2004. Massive horizontal transfer of mitochondrial genes from diverse land plant donors to the basal angiosperm *Amborella*. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51), pp.17747-17752. doi:10.1073/pnas.0408336102.

Blount, Z.D., Barrick, J.E., Davidson, C.J. and Lenski, R.E. (2012). Genomic analysis of a key innovation in an experimental *Escherichia coli* population. *Nature*, 489(7417), pp.513–518. doi:10.1038/nature11514.

Bofkin, L. and Goldman, N., 2006. Variation in evolutionary processes at different codon positions. *Molecular Biology and Evolution*, 24(2), pp.513-521. doi.org/10.1093/molbev/msl178.

Boothby, T.C., Tenlen, J.R., Smith, F.W., Wang, J.R., Patanella, K.A., Nishimura E.O, Tintori, S.C., Li, Q., Jones, C.D., Yandell, M., Messina, D.N., Glasscock, J. and Goldstein, B., 2015. Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proceedings of the National Academy of Sciences of the United States of America*, 112(52), pp.15976-15981. doi:10.1073/pnas.1510461112.

Boucher, Y., Douady, C.J., Papke, R.T., Walsh, D.A., Boudreau, M.R., Nesbø, C.L., Case, R.J. and Doolittle, W.F, 2003. Lateral gene transfer and the origins of prokaryotic groups. *Annual Review of Genetics*, 37(1), pp.283-328. doi:10.1146/annurev.genet.37.050503.084247.

Bowman, J.L., Kohchi, T., Yamato, K.T., Jenkins, J., Shu, S., Ishizaki, K., Yamaoka, S., Nishihama, R., Nakamura, Y., Berger, F., Adam, C., Aki, S.S., Althoff, F., Araki, T., Arteaga-Vazquez, M.A., Balasubrmanian, S., Barry, K., Bauer, D., Boehm, C.R., Briginshaw, L., Caballero-Perez, J., Catarino, B., Chen, F., Chiyoda, S., Chovatia, M., Davies, K.M., Delmans, M., Demura, T., Dierschke, T., Dolan, L., Dorantes-Acosta, A.E., Eklund, D.M., Florent, S.N., Flores-Sandoval, E., Fujiyama, A., Fukuzawa, H., Galik, B., Grimanelli, D., Grimwood, J., Grossniklaus, U., Hamada, T., Haseloff, J., Hetherington, A.J., Higo, A., Hirakawa, Y., Hundley, H.N., Ikeda, Y., Inoue, K., Inoue, S.I., Ishida, S., Jia, Q., Kakita, M., Kanazawa, T., Kawai, Y., Kawashima, T., Kennedy, M., Kinose, K., Kinoshita, T., Kohara, Y., Koide, E., Komatsu, K., Kopischke, S., Kubo, M., Kyozuka, J., Lagercrantz, U., Lin, S.S., Lindquist, E., Lipzen, A.M., Lu, C.W., De Luna, E., Martienssen, R.A., Minamino, N., Mizutani, M., Mochizuki, N., Monte, I., Mosher, R., Nagasaki, H., Nakagami, H., Naramoto, S., Nishitani, K., Ohtani, M., Okamoto, T., Okumura, M., Phillips, J., Pollak, B., Reinders, A., Rövekamp, M., Sano, R., Sawa, S., Schmid, M.W., Shirakawa, M., Solano, R., Spunde, A., Suetsugu, N., Sugano, S., Sugiyama, A., Sun, R., Suzuki, Y., Takenaka, M., Takezawa, D., Tomogane, H., Tsuzuki, M., Ueda, T., Umeda, M., Ward, J.M., Watanabe, Y., Yazaki, K., Yokoyama, R., Yoshitake, Y., Yotsui, I., Zachgo, S. and Schmutz J., 2017. Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell* 171(2), pp. 287-304. doi:10.1016/j.cell.2017.09.030.

Boyle, E.A., Li, Y.I. and Pritchard, J.K., 2017. An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7), pp.1177-1186. doi:10.1016/j.cell.2017.05.038.

Broad Institute. 2019. Picard Toolkit. GitHub repository. http://broadinstitute.github.io/picard/ [accessed 16 September 2020]

Cai, L., Arnold, B.J, Xi, Z., Khost, D.E., Patel, N., Hartmann, CB., Manickam, S., Sasirat, S., Nikolov, L.A., Mathews, S., Sackton, T.B. and Davis, C.C., 2021. Deeply altered genome architecture in the endoparasitic flowering plant *Sapria himalayana* Griff. (Rafflesiaceae). *Current Biology*, 31(5), pp.1002-1011.e9. doi:10.1016/j.cub.2020.12.045.

Cannarozzi, G., Plaza-Wüthrich, S., Esfeld, K., Larti, S., Wilson, Y. S., Girma, D., de Castro, E., Chanyalew, S., Blösch, R., Farinelli, L., Lyons, E., Schneider, M., Falquet, L., Kuhlemeier, C., Assefa, K. and Tadele, Z., 2014. Genome and transcriptome sequencing identifies breeding targets in the orphan crop tef (*Eragrostis tef*). *BMC Genomics* 15(1), pp. 1-21. doi:10.1186/1471-2164-15-581.

Chandrasekaran, C. and Betrán, E., 2008 Origins of new genes and pseudogenes. *Nature Education* 1(1), pp.181

Chang, B.S. and Campbell, D.L., 2000. Bias in phylogenetic reconstruction of vertebrate rhodopsin sequences. *Molecular Biology and Evolution* 17(8), pp. 1220-31. doi:10.1093/oxfordjournals.molbev.a026405.

Chen, F., Dong, W., Zhang, J., Guo, X., Chen, J., Wang, Z., Lin, Z., Tang, H. and Zhang, L., 2018. The sequenced angiosperm genomes and genome databases. *Frontiers in Plant Science,* 9(418), pp. 1-18. doi:10.3389/fpls.2018.00418.

Chen, R., Huangfu L., Lu, Y., Fang, H., Xu, Y., Li, P., Zhou, Y., Xu, C., Huang, J. and Yang, Z., 2021. Adaptive innovation of green plants by horizontal gene transfer. *Biotechnology Advances*, 46(107671). doi:10.1016/j.biotechadv.2020.107671.

Cheng, S., Xian, W., Fu, Y., Marin, B., Keller, J., Wu, T, Sun, W., Li, X., Xu, Y., Zhang, Y. and Wittek, S., 2019. Genomes of subaerial Zygnematophyceae provide insights into land plant evolution. *Cell* 179 (5), pp. 1057-67. doi:10.1016/j.cell.2019.10.019.

Christin P.A., Edwards E.J., Besnard G., Boxall S.F., Gregory R., Kellogg E.A., Hartwell J. and Osborne C.P., 2012a. Adaptive evolution of C_4 photosynthesis through recurrent lateral gene transfer. *Current Biology* 22(5), pp. 445-449. doi:10.1016/j.cub.2012.01.054.

Christin P.A., Besnard G., Edwards E.J. and Salamin N., 2012b. Effect of genetic convergence on phylogenetic inference. *Molecular Phylogenetics and Evolution* 62(3), pp. 921-927. doi:10.1016/j.ympev.2011.12.002.

Christin, P.A., Wallace, M.J., Clayton, H., Edwards, E.J., Furbank, R.T., Hattersley, P.W., Sage, R.F., Macfarlane, T.D. and Ludwig, M., 2012c. Multiple photosynthetic transitions, polyploidy, and lateral gene transfer in the grass subtribe Neurachninae. *Journal of Experimental Botany*, 63(17), pp.6297–6308. doi:10.1093/jxb/ers282.

Christin P.A., Spriggs E., Osborne C.P., Strömberg C.A., Salamin N. and Edwards E.J., 2014. Molecular dating, evolutionary rates, and the age of the grasses. *Systematic Biology* 63(2), pp. 153-165. doi:10.1093/sysbio/syt072.

Christin, P.A. and Osborne, C.P., 2014. The evolutionary ecology of C_4 plants. *New Phytologist*, 204(4), pp.765-781. doi:10.1111/nph.13033.

Christin, P.A., Besnard, G., Samaritani, E., Duvall, M.R., Hodkinson, T.R., Savolainen, V. and Salamin N., 2008. Oligocene CO_2 decline promoted C_4 photosynthesis in grasses. *Current Biology*, 18(1), pp.37-43. doi:10.1016/j.cub.2007.11.058.

Christin, P.A., Petitpierre, B., Salamin, N., Büchi, L. and Besnard, G., 2009. Evolution of C_4 phosphoenolpyruvate carboxykinase in grasses, from genotype to phenotype. *Molecular Biology and Evolution*, 26(2), pp.357-365. doi:10.1093/molbev/msn255.

Christin, P.A., Salamin, N., Savolainen, V., Duvall, M.R. and Besnard, G., 2007. C_4 Photosynthesis evolved in grasses via parallel adaptive genetic changes. *Current Biology*, 17(14), pp.1241-1247. doi:10.1016/j.cub.2007.06.036

Clark, J., Hidalgo, O., Pellicer, J., Liu, H., Marquardt, J., Robert, Y., Christenhusz, M., Zhang, S., Gibby, M., Leitch, I.J. and Schneider, H., 2016. Genome evolution of ferns: evidence for relative stasis of genome size across the fern phylogeny. *New Phytologist*, 210(3), pp.1072-1082. doi:10.1111/nph.13833.

Clayton, W.D., Vorontsova M.S., Harman, K.T. and Williamson H., 2016. GrassBase - The Online World Grass Flora. https://www.kew.org/data/grassbase/ [accessed 15 July 2019].

Cong, Y., Chan, Y. and Ragan, M.A., 2016. A novel alignment-free method for detection of lateral genetic transfer based on TF-IDF. *Scientific Reports*, 6(1). doi:10.1038/srep30308.

Cordero O.X. and Hogeweg, P., 2009. The impact of long-distance horizontal gene transfer on prokaryotic genome size. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51), pp.21748-21753. doi:10.1073/pnas.0907584106.

Cullis, C.A., Vorster, B.J., Van Der Vyver, C. and Kunert, K.J., 2008. Transfer of genetic material between the chloroplast and nucleus: how is it related to stress in plants?. *Annals of Botany*, 103(4), pp.625-633. doi:10.1093/aob/mcn173.

Danchin, E.G.J., 2016. Lateral gene transfer in eukaryotes: tip of the iceberg or of the ice cube?. *BMC Biology* 14(101), pp.1-3. doi:10.1186/s12915-016-0330-x.

Davies, J., 1995. Vicious circles: looking back on resistance plasmids. *Genetics*, 139(4), pp.1465-1468. doi:10.1093/genetics/139.4.1465.

Davis, C.C., Anderson, W.R. and Wurdack, K.J, 2005. Gene transfer from a parasitic flowering plant to a fern. *Proceedings of the Royal Society B: Biological Sciences*, 272(1578), pp.2237-2242. doi:10.1098/rspb.2005.3226.

de Felipe, K.S., Pampou, S., Jovanovic, O.S., Pericone, C.D., Ye, S.F., Kalachikov, S. and Shuman, H.A. (2005). Evidence for acquisition of legionella type IV secretion substrates via interdomain horizontal gene transfer. *Journal of Bacteriology,* 187(22), pp.7716–7726. doi:10.1128/JB.187.22.7716-7726.2005.

Degnan, J.H. and Rosenberg, N.A., 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. T*rends in Ecology & Evolution*, 24(6), pp.332-340. doi:10.1016/j.tree.2009.01.009.

Deng, C., Cheng, C.H., Ye, H., He, X. and Chen, L., 2010. Evolution of an antifreeze protein by neofunctionalization under escape from adaptive conflict. *Proceedings of the National Academy of Sciences of the United States of America*, 107(50), pp.21593-21598. doi:10.1073/pnas.1007883107.

Deusch, O., Landan, G., Roettger, M., Gruenheit, N., Kowallik, K.V., Allen, J.F., Martin, W. and Dagan, T., 2008. Genes of cyanobacterial origin in plant nuclear genomes point to a heterocyst-forming plastid ancestor. *Molecular Biology and Evolution*, 25(4), pp.748-761. doi: 10.1093/molbev/msn022.

Dierckxsens, N., Mardulyn, P. and Smits, G., 2016. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, 45(4) p.g 18. doi:10.1093/nar/gkw955.

Doležel, J., Greilhuber, J. and Suda, J., 2007. Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols*, 2(9), pp.2233-2244. doi:10.1038/nprot.2007.310.

Domazet-Lošo, M. and Haubold, B., 2011. Alignment-free detection of local similarity among viral and bacterial genomes. *Bioinformatics*, 27(11), pp.1466-1472. doi:10.1093/bioinformatics/btr176.

Doolittle, W.F. 1999. Lateral genomics. *Trends in Biochemical Sciences* 9(12), pp. M5-M8.

Dunning, L.T., Olofsson, J.K., Parisod, C., Choudhury, R.R., Moreno-Villena, J.J., Yang, Y., Dinora, J., Quick, W.P., Park, M., Bennetzen, J.L., Besnard, B. Nosil, P., Colin, C.P and Christin P,A,. 2019. Lateral transfers of large DNA fragments spread functional genes among grasses. *Proceedings of the National Academy of Sciences of the United States of America* 116 (10), pp. 4416-4425. doi:10.1073/pnas.1810031116.

Dunning, L.T., Lundgren, M.R., Moreno-Villena, J.J., Namaganda, M., Edwards, E.J., Nosil, P., Osborne, C.P. and Christin, P.A., 2017. Introgression and repeated co-option facilitated the recurrent emergence of C_4 photosynthesis among close relatives. *Evolution*, 71(6), pp.1541-1555. doi:10.1111/evo.13250.

Dunning, L.T., Olofsson, J.K., Parisod, C., Choudhury, R.R., Moreno-Villena, J.J., Yang, Y., Dionora, J., Quick, W. P., Park, M., Bennetzen, J.L., Besnard, G., Nosil, P., Osborne, C.P. and Christin, P.A,

2019. Lateral transfers of large DNA fragments spread functional genes among grasses. *Proceedings of the National Academy of Sciences of the United States of America*, 116(10), pp.4416-4425. doi:10.1073/pnas.1810031116.

El Baidouri, M., Carpentier, M.C., Cooke, R., Gao, D., Lasserre, E., Llauro, C., Mirouze, M., Picault, N., Jackson, S.A. and Panaud, O., 2014. Widespread and frequent horizontal transfers of transposable elements in plants. *Genome Research*, 24(5), pp.831-838. doi:10.1101/gr.164400.113.

Ellison, A.M. and Gotelli, N.J., 2009. Energetics and the evolution of carnivorous plants—Darwin's ‘most wonderful plants in the world’. *Journal of Experimental Botany*, 60(1), pp.19-42. doi:10.1093/jxb/ern179.

Emms, D.M. and Kelly S., 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves ortholog inference accuracy. *Genome Biology* 16(157), pp. 1-14.

Felsenstein, J., 1988. Phylogenies and quantitative characteristics. *Annual Review of Ecology and Systematics*, 19(1), pp.445-471.

Fick, S.E. and Hijmans, R.J., 2017. WorldClim 2: new 1 km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* 37(12), pp. 4302-4315. doi:10.1002/joc.5086.

Freeman, V., 1951. Studies on the virulence of bacteriophage-infected strains of *Corynebacterium diphtheriae*. *Journal of Bacteriology*, 61(6), pp.675-688.

Freschi, L., Vincent, A.T., Jeukens, J., Emond-Rheault, J.G., Kukavica-Ibrulj, I., Dupont, M.J., Charette, S.J., Boyle, B. and Levesque, R.C., 2018. The *Pseudomonas aeruginosa* pan-genome provides new insights on its population structure, horizontal gene transfer, and pathogenicity. *Genome Biology and Evolution* 11(1), pp.109-120. doi:10.1093/gbe/evy259.

Fuglie, K., Peters, M. and Burkart, S., 2021. The extent and economic significance of cultivated forage crops in developing countries. *Frontier in Sustainable Food Systems* 5, 712136. doi:10.3389/fsufs.2021.712136.

Gao, C., Ren, X., Mason, A.S., Liu, H., Xiao, M., Li, J. and Fu, D., 2014. Horizontal gene transfer in plants. *Functional & Integrative Genomics* 14(1), pp. 23-29. doi:10.1007/s10142-013-0345-0.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B.M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W.L., Chen, L., Cooper, B., Park, S., Wood, T.C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R.M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A. and Briggs, S., 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296(5565), pp. 92-100. doi:10.1126/science.1068275.

Graham, B. and Bormann, F., 1966. Natural root grafts. *The Botanical Review*, 32(3), pp.255-292.

Grass Phylogeny Working Group II. 2012. New grass phylogeny resolves deep evolutionary relationships and discovers C_4 origins. *New Phytologist* 193(2), pp.304– 312. doi:10.1111/j.1469-8137.2011.03972.x.

Griffith, F., 1928. The Significance of pneumococcal types. *Journal of Hygiene*, 27(2), pp.113-159.

Guindon, S. and Gascuel, O., 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* 52(5), pp.696-704. doi:10.1080/10635150390235520.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W. and Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3), pp.307-321. doi:10.1093/sysbio/syq010.

Guo, L., Qiu, J., Ye, C., Jin, G., Mao, L., Zhang, H., Yang, X., Peng, Q., Wang, Y., Jia, L., Lin, Z., Li, G., Fu, F., Liu, C., Chen, L., Shen, E., Wang, W., Chu, Q., Wu, D., Wu, S., Xia, C., Zhang, Y., Zhou, X., Wang, L., Wu, L., Song, W., Wang, Y., Shu, Q., Aoki, D., Yumoto, E., Yokota, T., Miyamoto, K., Okada, K., Kim, D.S., Cai, D., Zhang, C., Lou, Y., Qian, Q., Yamaguchi, H., Yamane, H., Kong, C.H., Timko, M.P., Bai, L. and Fan, L., 2017. *Echinochloa crus-galli* genome analysis provides insight into its adaptation and invasiveness as a weed. *Nature Communications*, 8(1):1031, pp.1-10. doi:10.1038/s41467-017-01067-5.

Ha, M., Kim, E. and Chen, Z.J., 2009. Duplicate genes increase expression diversity in closely related species and allopolyploids. *Proceedings of the National Academy of Sciences of the United States of America*, 106(7), pp.2295-2300. doi:10.1073/pnas.0807350106.

Harris, I.P.D.J., Jones, P.D., Osborn, T.J. and Lister, D.H., 2014. Updated high resolution grids of monthly climatic observations–the CRU TS3. 10 Dataset. *International Journal of Climatology* 34(3), pp.623-642. doi:10.1002/joc.3711.

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

Hertle, A.P., Haberl, B. and Bock, R., 2021. Horizontal genome transfer by cell-to-cell travel of whole organelles. *Science Advances* 7(1), pp. 1. doi:10.1126/sciadv.abd8215.

Hibdige, S.G.S., Raimondeau, P., Christin, P.A and Dunning, L.T., 2021. Widespread lateral gene transfer among grasses. *New Phytologist*, 230(6), pp.2474-2486. doi:10.1111/nph.17328.

Hu, Y., Wu, Q., Ma, S., Ma, T., Shan, L., Wang, X., Nie, Y., Ning, Z., Yan, L., Xiu, Y. and Wei, F., 2017. Comparative genomics reveals convergent evolution between the bamboo-eating giant and red pandas. *Proceedings of the National Academy of Sciences of the United States of America*, 114(5), pp.1081-1086. doi:10.1073/pnas.1613870114.

Huang, P., Studer, A.J., Schnable, J.C., Kellogg, E.A. and Brutnell, T.P., 2016. Cross species selection scans identify components of C_4 photosynthesis in the grasses. *Journal of Experimental Botany*, 68(2), pp.127-135. doi:10.1093/jxb/erw256.

Husnik, F. and McCutcheon, J.P., 2018. Functional horizontal gene transfer from bacteria to eukaryotes. *Nature Reviews Microbiology* 16(2018), pp.67-79.

Husnik, F., Nikoh, N., Koga, R., Ross, L., Duncan, R.P., Fujie, M., Tanaka, M., Satoh, N., Bachtrog, D., Wilson, A.C.C, von Dohlen, C.D., Fukatsu, T. and McCutcheon, J.P., 2013. Horizontal gene transfer from diverse bacteria to an insect genome enables a tripartite nested mealybug symbiosis. *Cell*, 153(7), pp.1567-1578. doi:10.1016/j.cell.2013.05.040.

International Barley Genome Sequencing Consortium. 2012. A physical, genetic and functional sequence assembly of the barley genome. *Nature*, 491(2012), pp.711-716. doi:10.1038/nature11543.

International Brachypodium Initiative. 2010. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature,* 463, pp.763-768. doi:10.1038/nature08747.

International Wheat Genome Sequencing Consortium. 2014. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science*, 345(6194), pp.1251788. doi:10.1126/science.1251788.

Kanai, R. and Edwards, G.E., 1999. The biochemistry of C_4 photosynthesis. C_4 *plant biology*, 49(1), p.87.

Katoh, K. and Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* 30(4), pp.772-780. doi:10.1093/molbev/mst010.

Kawahara, A.Y., Plotkin, D., Espeland, M., Meusemann, K., Toussaint, E.F.A., Donath, A., Gimnich, F., Frandsen, P.B., Zwick, A., Reis, M. dos, Barber, J.R., Peters, R.S., Liu, S., Zhou, X., Mayer, C., Podsiadlowski, L., Storer, C., Yack, J.E., Misof, B. and Breinholt, J.W. (2019). Phylogenomics reveals the evolutionary timing and pattern of butterflies and moths. Proceedings of the National Academy of Sciences, 116(45), pp.22657–22663. doi:10.1073/pnas.1907847116.

Keeling, P.J. and Palmer, J.D, 2008. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics*, 9(8), pp.605-618. doi:10.1038/nrg2386.

Kellogg, E.A. and Buell. C.R., 2009. Splendor in the grasses. *Plant Physiology* 149(1), pp.1-3. doi:10.1104/pp.104.900281.

Kim, S., Park, M., Yeom, S.I., Kim, Y.M., Lee, J.M., Lee, H.A., Seo, E., Choi, J., Cheong, K., Kim, K.T., Jung, K., Lee, G.-W., Oh, S.-K., Bae, C., Kim, S.B., Lee, H.Y., Kim, S.Y., Kim, M.S., Kang, B.C. and Jo, Y.D. (2014). Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. *Nature Genetics*, 46(3), pp.270–278. doi:10.1038/ng.2877.

Kleine, T., Maier, U.G. and Leister, D., 2009. DNA transfer from organelles to the nucleus: The idiosyncratic genetics of endosymbiosis. *Annual Review of Plant Biology*, 60(1), pp.115-138. doi:10.1146/annurev.arplant.043008.092119.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H. and Phillippy, A.M., 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, 27(5), pp.722-736.

Koutsovoulos, G., Kumar, S., Laetsch, D., Stevens, L., Daub, J., Conlon, C., Maroon, H., Thomas, F., Aboobaker, A. and Blaxter, M., 2016. No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. *Proceedings of the National Academy of Sciences of the United States of America*, 113(18), pp.5053-5058.

Kyndt, T., Quispe, D., Zhai, H., Jarret, R., Ghislain, M., Liu, Q., Gheysen, G. and Kreuze, J., 2015. The genome of cultivated sweet potato contains *Agrobacterium* T-DNAs with expressed genes: An example of a naturally transgenic food crop. *Proceedings of the National Academy of Sciences of the United States of America*, 112(18), pp.5844-5849. doi:10.1073/pnas.1419685112.

Lang, A.S., Zhaxybayeva, O. and Beatty, J.T., 2012. Gene transfer agents: phage-like elements of genetic exchange. *Nature Reviews Microbiology*, 10(7), pp.472-482. doi:10.1038/nrmicro2802.

Langmead, B. and Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9(4), pp.357-359. doi:10.1038/nmeth.1923.

Lederberg, J. and Tatum, E.L, 1946. Gene Recombination in *Escherichia Coli*. *Nature*, 158(4016), pp.558-558.

Lefort, V., Longueville, J. and Gascuel, O., 2017. SMS: Smart model selection in PhyML. *Molecular Biology and Evolution*, 34(9), pp.2422-2424.

Lehmann, C.E., Griffith, D.M., Simpson, K.J., Anderson, T.M., Archibald, S., Beerling, D.J., Bond, W.J., Denton, E., Edwards, E.J., Forrestel, E.J., Fox, D.L., Georges, D., Hoffman, W.A., Kluyver, T., Mucina, L., Pau, S., Salamin, N., Santini, B., Smith, M.D., Spriggs, E.L., Westley, R., Still, C.J. Stömberg, C.A.E. and Osborne, C.P., 2019. Functional diversification enabled grassy biomes to fill global climate space. *BioRxiv* 583625. doi:10.1101/583625.

Lerat, E., Daubin, V., Ochman, H. and Moran, N., 2005. Evolutionary origins of genomic repertoires in bacteria. *PLoS Biology*, 3(5), p.p130. doi:10.1371/journal.pbio.0030130.

Li. F.W., Villarreal, J.C., Kelly, S., Rothfels, C.J., Melkonian, M., Frangedakis, E., Ruhsam, M., Sigel, E.M., Der, J.P., Pittermann, J., Burge, D.O., 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, D.W., Crandall-Stotler, B.J., Shaw, A.J., Deyholos, M.K., Soltis, D.E., Graham, S.W., Windham, M.D., Langdale, J.A., Wong, G. K.S., Mathews, S. and Pryer, K.M., 2014. Horizontal transfer of an adaptive chimeric photoreceptor from bryophytes to ferns. *Proceedings of the National Academy of Sciences of the United States of America*, 111(18), pp.6672-6677. doi:10.1073/pnas.1319929111.

Li, F.W., Brouwer, P., Carretero-Paulet, L., Cheng, S., de Vries, J., Delaux, P.M., Eily, A., Koppers, N., Kuo, L., Li, Z., Simenc, M., Small, I., Wafula, E., Angarita, S., Barker, M., Bräutigam, A., dePamphilis, C., Gould, S., Hosmani, P., Huang, Y., Huettel, B., Kato, Y., Liu, X., Maere, S., McDowell, R., Mueller, L., Nierop, K.J., Rensing, S.A., Robison, T., Rothfels, C.J., Sigel, E.M., Song, Y., Timilsena, P.R., Van de Peer, Y., Wang, H., Wilhelmsson, P.K.I., Wolf, P.G., Xu, X., Der, J.P., Schluepmann, H., Wong, G.K.S. and Pryer, K.M., 2018. Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nature Plants*, 4(7), pp.460-472. doi:10.1038/s41477-018-0188-8.

Lindow, S.E., 2017. Horizontal gene transfer gone wild: promiscuity in a kiwifruit pathogen leads to resistance to chemical control. *Environmental Microbiology*, 19(4), pp.1363-1365.

Lovell, J.T., Jenkins, J., Lowry, D.B., Mamidi, S., Sreedasyam, A., Weng, X., Barry, K., Bonnette, J., Campitelli, B., Daum, C., Gordon, S.P., Gould, B.A., Khasanova, A., Lipzen, A., Macqueen, A., Palacio-Mejía, J.D., Plott, C., Shakirov, E.V., Shu, S., Yoshinaga, Y., Zane, M., Kudrna, D., Talag, J.D., Rokhsar, D., Grimwood, J., Schmutz, J. and Juenger, T.E., 2018. The genomic landscape of molecular responses to natural drought stress in *Panicum hallii*. *Nature Communications,* 9, pp. 5213. doi:10.1038/s41467-018-07669-x.

Lyu, J., Huang, L., Zhang, S., Zhang, Y., He, W., Zeng, P., Zeng, Y., Huang, G., Zhang, J., Ning, M., Bao, Y., Zhao, S., Fu, Q., Wade, L.G., Chen, H., Wang, W. and Hu, F., 2020. Neo-functionalization of a teosinte branched 1 homologue mediates adaptations of upland rice. *Nature Communications,* 11:725(2020). doi:10.1038/s41467-019-14264-1.

Ma, J., Wang, S., Zhu, X., Sun, G., Chang, G., Li, L., Hu, X., Zhang, S., Zhou, Y., Song, C.P. and Huang, J., 2022. Major episodes of horizontal gene transfer drove the evolution of land plants. *Molecular Plant*, 15(5), pp.857-871.

Maddison, W.P. (1997). Gene trees in species trees. *Systematic Biology*, 46(3), pp.523–536. doi:10.1093/sysbio/46.3.523.

Mahelka, V., Krak, K., Kopecký, D., Fehrer, J., Šafář, J., Bartoš, J. and Blattner, F.R., 2017. Multiple horizontal transfers of nuclear ribosomal genes between phylogenetically distinct grass lineages. *Proceedings of the National Academy of Sciences of the United States of America*, 114(7), pp.1726-1731. doi:10.1073/pnas.1613375114.

Mahelka, V., Krak, K., Fehrer, J., Caklová, P., Nagy Nejedlá, M., Čegan, R., Kopecký, D. and Šafář, J. (2021). A *Panicum* -derived chromosomal segment captured by *Hordeum* a few million years ago preserves a set of stress-related genes. *The Plant Journal*, 105(5), pp.1141–1164. doi:10.1111/tpj.15167.

Martin, W.F., 2017. Too much eukaryote LGT. *BioEssays*, 39(12), p.1700115. doi:10.1002/bies.201700115.

Maumus, F., Epert, A., Nogué, F. and Blanc, G. 2014. Plant genomes enclose footprints of past infections by giant virus relatives. *Nature Communications,* 5, 4268 doi:10.1038/ncomms5268.

McDaniel, L.D., Young, E., Delaney, J., Ruhnau, F., Ritchie, K.B. and Paul, J.H., 2010. High frequency of horizontal gene transfer in the oceans. *Science*, 330(6000), pp.50-50. doi:10.1126/science.1192243.

Medini, D., Donati, C., Tettelin, H., Masignani, V. and Rappuoli, R., 2005. The microbial pan-genome. *Current Opinion in Genetics & Development*, 15(6), pp.589-594. doi:10.1016/j.gde.2005.09.006.

Metzger, M.J., Paynter, A.N., Siddall, M.E. and Goff, S.P., 2018. Horizontal transfer of retrotransposons between bivalves and other aquatic species of multiple phyla. *Proceedings of the National Academy of Sciences of the United States of America*, 115(18), pp.E4227-E4235. doi:10.1073/pnas.1717227115.

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

Moreno-Villena, J.J., Dunning, L.T., Osborne, C.P. and Christin, P.A., 2017. Highly expressed genes are preferentially co-opted for C_4 photosynthesis. *Molecular Biology and Evolution* 35(1), pp. 94-106. doi:10.1093/molbev/msx269.

Mousseau, T. and Roff, D., 1987. Natural selection and the heritability of fitness components. *Heredity*, 59(2), pp.181-197.

Mower, J.P., Stefanović, S., Young, G.J. and Palmer, J.D., 2004. Gene transfer from parasitic to host plants. *Nature*, 432(7014), pp.165-166.

Nakamura, Y., Itoh, T., Matsuda, H. and Gojobori, T., 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nature Genetics*, 36(7), pp.760-766. doi:10.1038/ng1381.

Naturalearthdata.com. 2022. Natural Earth - Free vector and raster map data at 1:10m, 1:50m, and 1:110m scales. Available at: <https://www.naturalearthdata.com/> [Accessed 22 May 2022].

Nichols, R., 2001. Gene trees and species trees are not the same. *Trends in Ecology & Evolution*, 16(7), pp.358-364.

Nikoh, N. and Nakabachi, A., 2009. Aphids acquired symbiotic genes via lateral gene transfer. *BMC Biology*, 7, 12. doi:10.1186/1741-7007-7-12.

Nováková, E. and Moran, N.A., 2011. Diversification of genes for carotenoid biosynthesis in aphids following an ancient transfer from a fungus. *Molecular Biology and Evolution*, 29(1), pp.313-323.

Nowack, E.C.M, Vogel, H., Groth, M., Grossman, A.R., Melkonian, M. and Glockner, G., 2010. Endosymbiotic gene transfer and transcriptional regulation of transferred genes in *Paulinella chromatophora*. *Molecular Biology and Evolution*, 28(1), pp.407-422. doi:10.1093/molbev/msq209.

Ochman, H., Lawrence, J.G. and Groisman, E.A., 2000. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784), pp.299-304. doi:10.1038/35012500.

Olofsson, J.K., Dunning, L.T., Lundgren, M.R., Barton, H.J., Thompson, J., Cuff, N., Ariyarathne, M., Yakandawala, D., Sotelo, G., Zeng, K., Osborne, C.P., Patrik, P. and Christin, P.A.,. 2019. Population-specific selection on standing variation generated by lateral gene transfers in a grass. *Current Biology* 29(22), pp.3921-7. doi:10.1016/j.cub.2019.09.023.

Olofsson, J.K., Bianconi, M., Besnard, G., Dunning, L.T., Lundgren, M., Holota, H.R., Vorontsova, M.S., Hidalgo, O., Leitch, I.J., Nosil, P., Osborne, C.P. and Christin, P.A., 2016. Genome biogeography reveals the intraspecific spread of adaptive mutations for a complex trait. *Molecular Ecology*, 25(24), pp.6107-6123. doi:10.1111/mec.13914.

O'Mara, F.P., 2012. The role of grasslands in food security and climate change. *Annals of Botany*, 110(6), pp.1263-1270.

One Thousand Plant Transcriptomes Initiative., 2019 One thousand plant transcriptomes and the phylogenomics of green plants. *Nature* 574, pp. 679–685. https://doi.org/10.1038/s41586-019-1693-2

Orme, D., Freckleton, R.P., Thomas, G.H., Petzoldt, T. and Fritz, S., 2013. The caper package: comparative analysis of phylogenetics and evolution in R. *R package version 5.2*(2013),pp.1-36.

Pamilo, P. and Nei, M., 1988. Relationships between gene trees and species trees. *Molecular Biology and Evolution*, 5(5), pp.568-283.

Park, M., Christin, P.A. and Bennetzen, J.L., 2021a. Sample sequence analysis uncovers recurrent horizontal transfers of transposable elements among grasses. *Molecular Biology and Evolution*, 38(9), pp.3664-3675. doi:10.1093/molbev/msab133.

Park, M., Jo, S., Kwon, J.-K., Park, J., Ahn, J.H., Kim, S., Lee, Y.-H., Yang, T.-J., Hur, C.-G., Kang, B.-C., Kim, B.-D. and Choi, D., 2011a. Comparative analysis of pepper and tomato reveals euchromatin expansion of pepper genome caused by differential accumulation of Ty3/Gypsy-like elements. *BMC Genomics*, 12(1). doi:10.1186/1471-2164-12-85.

Park, M., Park, J., Kim, S., Kwon, J.-K., Park, H.M., Bae, I.H., Yang, T.-J., Lee, Y.-H., Kang, B.-C. and Choi, D., 2011b. Evolution of the large genome in *Capsicum annuum* occurred through accumulation of single-type long terminal repeat retrotransposons and their derivatives. *The Plant Journal*, 69(6), pp.1018–1029. doi:10.1111/j.1365-313x.2011.04851.x.

Park, M., Sarkhosh, A., Tsolova, V. and El-Sharkawy, I., 2021b. Horizontal transfer of LTR retrotransposons contributes to the genome diversity of *Vitis*. *International Journal of Molecular Sciences*, 22(19), p.10446. doi:10.3390/ijms221910446.

Paterson, A.H., Bowers, J.E., 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, A.K., Chapman, J., Feltus, F.A.,Gowik, U., Grigoriev, I.V., Lyons, E., Maher, C.A., Martis, M., Narechania, A., Otillar, R.P., Penning, B.W., Salamov, A.A., Wang, Y., Zhang, L., Carpita, N.C., Freeling, M., Gingle, A.R., Hash, C.T., Keller, B., Klein, P., Kresovich, S., McCann, M.C., Ming, R., Peterson, D.G., Rahman, M.B., Ware, D., Westhoff, P., Mayer, K.F.X., Messing, J. and Rokhsar, D.S., 2009. The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, 457, 551-556(2009). doi:10.1038/nature07723.

Peccoud, J., Loiseau, V., Cordaux, R. and Gilbert, C., 2017. Massive horizontal transfer of transposable elements in insects. *Proceedings of the National Academy of Sciences of the United States of America* 114(18), pp.4721-4726. doi:10.1073/pnas.1621178114.

Pellicer, J. and Leitch, I.J., 2020. The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *New Phytologist* 226(2), pp.301-305. doi:10.1111/nph.16261.

Phansopa, C., Dunning, L.T., Reid, J.D. and Christin, P.A, 2020. Lateral gene transfer acts as an evolutionary shortcut to efficient C_4 biochemistry. *Molecular Biology and Evolution*, 37(11), pp.3094-3104. doi:10.1093/molbev/msaa143.

Plants of the World Online. 2022. *Alloteropsis semialata* (R.Br.) Hitchc. | Plants of the World Online | Kew Science. Available at: <https://powo.science.kew.org/taxon/urn:lsid:ipni.org:names:9363-2> [Accessed 22 May 2022].

Popa, O. and Dagan, T., 2011. Trends and barriers to lateral gene transfer in prokaryotes. *Current Opinion in Microbiology* 14(5), pp.615-623. doi:10.1016/j.mib.2011.07.027.

Prendergast, H., Hattersley, P. and Stone, N., 1987. New structural/biochemical associations in leaf blades of C_4 grasses (Poaceae). *Functional Plant Biology*, 14(4), p.403.

Prentice, H.C., Li, Y., Lönn, M., Tunlid, A. and Ghatnekar, L., 2015. A horizontally transferred nuclear gene is associated with microhabitat variation in a natural plant population. P*roceedings of the Royal Society B: Biological Sciences* 282(1821), pp.20152453. doi:10.1098/rspb.2015.2453.

Quinlan, A.R. and Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26(6), pp.841-842. doi:10.1093/bioinformatics/btq033.

Rambaut, A., Drummond, A.J., Xie, D., Baele, G. and Suchard, M.A, 2018. Posterior summarization in bayesian phylogenetics using tracer 1.7. *Systematic Biology*, 67(5), pp.901-904. doi:10.1093/sysbio/syy032.

Rastogi, S. and Liberles, D.A., 2005. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evolutionary Biology.* 5(28). doi: 10.1186/1471-2148-5-28

Reeves, G., Tripathi, A., Singh, P., Jones, M.R., Nanda, A.K., Musseau, C., Craze, M., Bowden, S., Walker, J.F., Bentley, A.W., Melnyk, C.W. and Hibberd, J.M, 2021. Monocotyledonous plants graft at the embryonic root–shoot interface. *Nature*, 602(7896), pp.280-286. doi:10.1038/s41586-021-04247-y.

Reneker, J., Lyons, E., Conant, G.C., Pires, J.C., Freeling, M., Shyu, C.R. and Korkin, D., 2012. Long identical multispecies elements in plant and animal genomes. *Proceedings of the National Academy of Sciences of the United States of America*, 109(19). pp. E1183-E1191.

Reyes-Prieto, A., Hackett, J.D., Soares, M.B., Bonaldo, M.F. and Bhattacharya, D., 2006. Cyanobacterial contribution to algal nuclear genomes is primarily limited to plastid functions. *Current Biology*, 16(23), pp.2320-2325. doi:10.1016/j.cub.2006.09.063.

Reynolds, H.T., Vijayakumar, V., Gluck-Thaler, E., Korotkin, H.B., Matheny, P. and Slot, J.C., 2018. Horizontal gene cluster transfer increased hallucinogenic mushroom diversity. *Evolution Letters*, 2(2), pp.88-101. doi:10.1002/evl3.42.

Rice, A., Glick, L., Abadi, S., Einhorn, M., Kopelman, N.M., Salman-Minkov, A., Mayzel, J., Chay, O. and Mayrose, I., 2015, The Chromosome Counts Database (CCDB) – a community resource of plant chromosome numbers. *New Phytologist*, 206, pp.19-26. https://doi.org/10.1111/nph.13191

Richardson, A.O. and Palmer, J.D., 2007. Horizontal gene transfer in plants. *Journal of Experimental Botany* 58(1), pp.1-9. doi:10.1093/jxb/erl148.

Saarela, J.M., Burke, S.V., Wysocki, W.P., Barrett, M.D., Clark, L.G., Craine, J.M., Peterson, P.M., Soreng, R.J., Vorontsova, M.S. and Duvall, M.R., 2018. A 250 plastome phylogeny of the grass family (Poaceae): topological support under different data partitions. *PeerJ*, 6, p.e4299.

Sage, R., 2003. The evolution of C_4 photosynthesis. *New Phytologist*, 161(2), pp.341-370. doi:10.1111/j.1469-8137.2004.00974.x.

Scally, A., Dutheil, J.Y., Hillier, L.W., Jordan, G.E., Goodhead, I., Herrero, J., Hobolth, A., Lappalainen, T., Mailund, T., Marques-Bonet, T., McCarthy, S., Montgomery, S.H., Schwalie, P.C., Tang, Y.A., Ward, M.C., Xue, Y., Yngvadottir, B., Alkan, C., Andersen, L.N. and Ayub, Q. (2012). Insights into hominid evolution from the gorilla genome sequence. *Nature*, 483(7388), pp.169–175. doi:10.1038/nature10842.

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., Graves, T.A., Minx, P., Reily, A.D., Courtney, L., Kruchowski, S.S., Tomlinson, C., Strong, C., Delehaunty, K., Fronick, C., Courtney, B. and Rock, S.M., 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science*, 326(5956), pp.1112–1115. doi:10.1126/science.1178534.

Schönknecht, G., Weber, A.P. and Lercher, M.J., 2014. Horizontal gene acquisitions by eukaryotes as drivers of adaptive evolution. *Bioessays* 36(1), pp.9-20.

Schwab, I., 2017. The evolution of eyes: major steps. The Keeler lecture 2017: centenary of Keeler Ltd. *Eye*, 32(2), pp.302-313. doi:10.1002/bies.201300095.

Shantz, H.L., 1954. The place of grasslands in the earth's Ccover. *Ecology*, 35(2), pp.143-145.

Shen, Z., Dong, X.M., Gao, Z.F., Chao, Q. and Wang, B.C., 2017. Phylogenic and phosphorylation regulation difference of phosphoenolpyruvate carboxykinase of C_3 and C_4 plants. *Journal of Plant Physiology*, 213, pp.16–22. doi:10.1016/j.jplph.2017.02.008.

Shi, H., Kichaev, G. and Pasaniuc, B., 2016. Contrasting the genetic architecture of 30 complex traits from summary association data. *The American Journal of Human Genetics*, 99(1), pp.139-153. doi:10.1016/j.ajhg.2016.05.013.

Shimodaira, H. and Hasegawa, M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17(12), pp.1246-1247. doi:10.1093/bioinformatics/17.12.1246.

Sieber, K.B., Bromley, R.E. and Dunning Hotopp, J. C., 2017. Lateral gene transfer between prokaryotes and eukaryotes. *Experimental Cell Research*, 358(2), pp.421-426.

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V. and Zdobnov, E.M, 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19), pp.3210-3212. doi:10.1093/bioinformatics/btv351.

Martin, S.H., Van Belleghem, S.M., 2017. Exploring evolutionary relationships across the genome using topology weighting, Genetics, 206 (1), pp. 429–438, https://doi.org/10.1534/genetics.116.194720

Skippington, E. and Ragan, M.A., 2012. Phylogeny rather than ecology or lifestyle biases the construction of *Escherichia coli–Shigella* genetic exchange communities. *Open Biology*, 2(9), pp.120112. doi:10.1098/rsob.120112.

Smit, A.F.A., Hubley, R. and Green, P., 2013. RepeatMasker Open-4.0. <http://www.repeatmasker.org>.

Som, A., 2015. Causes, consequences and solutions of phylogenetic incongruence. *Briefings in Bioinformatics*, 16(3), pp.536-548.

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

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

Stamatakis A., 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), pp.1312-1313. doi:10.1093/bioinformatics/btu033.

Stegemann, S. and Bock, R., 2009. Exchange of genetic material between cells in plant tissue grafts. *Science*, 324(5927), pp.649-651.

Stegemann, S., Keuthe, M., Greiner, S. and Bock, R., 2012. Horizontal transfer of chloroplast genomes between plant species. *Proceedings of the National Academy of Sciences of the United States of America*, 109(7), pp.2434-2438.

Stein, J.C., Yu, Y., Copetti, D., Zwickl, D.J., Zhang, L., Zhang, C., Chougule, K., Gao, D., Iwata, A., Goicoechea, J.L. and Wei, S., 2018. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nature Genetics* 50(2), pp.285-296. doi:10.1038/s41588-018-0040-0.

Studer, A.J., Schnable, J.C., Weissmann, S., Kolbe, A.R., McKain, M.R., Shao, Y., Cousins, A.B., Kellogg, E.A. and Brutnell, T.P., 2016. The draft genome of the C_3 panicoid grass species *Dichanthelium oligosanthes*. *Genome Biology*, 17(1). doi:10.1186/s13059-016-1080-3.

Sun, D., Jeannot, K., Xiao, Y. and Knapp, C.W., 2019. Editorial: Horizontal gene transfer mediated bacterial antibiotic resistance. *Frontiers Microbiology*. 10(2019):1933. doi:10.3389/fmicb.2019.01933.

Sun, G., Bai, S., Guan, Y., Wang, S., Wang, Q., Liu, Y., Liu, H., Goffinet, B., Zhou, Y., Paoletti, M., Hu, X., Haas, F.B., Fernandez-Pozo, N., Czyrt, A., Sun, H., Rensing, S.A. and Huang, J. (2020). Are fungi-derived genomic regions related to antagonism towards fungi in mosses? *New Phytologist*, 228(4), pp.1169–1175. doi:10.1111/nph.16776.

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

Szöllősi, G.J. and Daubin, V., 2012. Modeling gene family evolution and reconciling phylogenetic discord, *Methods in Molecular Biology*. 856(1), pp. 29-5. doi:10.1007/978-1-61779-585-5_2.

Tanaka, H., Hirakawa, H., Kosugi, S., Nakayama, S., Ono, A., Watanabe, A., Hashiguchi, M., Gondo, T., Ishigaki, G., Muguerza, M., Shimizu, K., Sawamura, N., Inoue, T., Shigeki, Y., Ohno, N., Tabata, S., Akashi, R. and Sato, S., 2016. Sequencing and comparative analyses of the genomes of zoysiagrasses. *DNA Research*, 23(2), pp.171–180. doi:10.1093/dnares/dsw006.

Tang, H., Bomhoff, M.D., Briones, E., Zhang, L., Schnable, J.C. and Lyons, E., 2015. SynFind: compiling syntenic regions across any set of genomes on demand. *Genome Biology and Evolution*, 7(12), pp.3286-3298. doi:10.1093/gbe/evv219.

Taniguchi, Y., Yamada, Y., Maruyama, O., Kuhara, S. and Ikeda, D. (2013). The purity measure for genomic regions leads to horizontally transferred genes *Journal of Bioinformatics and Computational Biology*, 11(06), p.1343002. doi:10.1142/s0219720013430026.

Timmis, J.N., Ayliffe, M.A., Huang, C.Y. and Martin, W., 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nature Reviews Genetics*, 5(2), pp.123-135. doi:10.1038/nrg1271.

Vallenback, P., Jaarola, M., Ghatnekar, L. and Bengtsson, B.O., 2008. Origin and timing of the horizontal transfer of a PgiC gene from Poa to *Festuca ovina*. *Molecular Phylogenetics and Evolution*, 46(3), pp.890-896. doi:10.1016/j.ympev.2007.11.031.

van der Weijde, T., Alvim Kamei, C.L., Torres, A.F., Vermerris, W., Dolstra, O., Visser, R.G.F. and Trindade, L.M., 2013. The potential of C_4 grasses for cellulosic biofuel production. *Frontiers in Plant Science*, 4(107), pp.1-18. doi:10.3389/fpls.2013.00107.

Van Etten, J. and Bhattacharya, D., 2020. Horizontal gene transfer in eukaryotes: not if, but how much? *Trends in Genetics,* 36(12), pp.915-925. doi:10.1016/j.tig.2020.08.006.

Van Buren, R., Bryant, D., Edger, P.P., Tang, H., Burgess, D., Challabathula, D., Spittle, K., Hall, R., Gu, J., Lyons, E., Freeling, M., Bartels, D., Ten Hallers, B., Hastie, A., Michael, T.P. and Mockler, T.C. (2015). Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature*, 527(7579), pp.508–511. doi:10.1038/nature15714.

Varshney, R.K., Shi, C., Thudi, M., Mariac, C., Wallace, J., Qi, P., Zhang, H., Zhao, Y., Wang, X., Rathore, A., Srivastava, R.K., Chitikineni, A., Fan, G., Bajaj, P., Punnuri, S., Gupta, S.K., Wang, H., Jiang, Y., Couderc, M. and Katta, M.A.V.S.K., 2017. Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. *Nature Biotechnology*, 35(10), pp.969–976. doi:10.1038/nbt.3943.

Vernikos, G., Medini, D., Riley, D.R. and Tettelin, H., 2015. Ten years of pan-genome analyses. *Current Opinion in Microbiology*, 23(1), pp.148-154. doi:10.1016/j.mib.2014.11.016.

Visscher, P.M., Hill, W.G. and Wray, N.W., 2008. Heritability in the genomics era — concepts and misconceptions. *Nature Reviews Genetics*, 9(4), pp.255-266.

Vogel, A., Schwacke, R., Denton, A.K., Usadel, B., Hollmann, J., Fischer, K., Bolger, A., Schmidt, M.H.W., Bolger, M.E., Gundlach, H., Mayer, K.F.X., Weiss-Schneeweiss, H., Temsch, E.M. and Krause, K., 2018. Footprints of parasitism in the genome of the parasitic flowering plant *Cuscuta campestris*. *Nature Communications*, 9(1), p.2515. doi:10.1038/s41467-018-04344-z.

von Wintersdorff, C.J.H., Penders, J., van Niekerk, J.M., Mills, N.D., Majumder, S., van Alphen, L.B., Savelkoul, P.H.M. and Wolffs, P.F.G., 2016. Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Frontiers in Microbiology*, 7(2016):173. doi:10.3389/fmicb.2016.00173.

Wadsworth, C.B., Arnold, B.J., Sater, M.R.A. and Grad, Y.H., 2018. Azithromycin resistance through interspecific acquisition of an epistasis-dependent efflux pump component and transcriptional regulator in neisseria gonorrhoeae. *mBio,* 9(4). doi:10.1128/mBio.01419-18.

Wagner, A., 2002. Selection and gene duplication: a view from the genome. *Genome Biology,* 3(5) pp 1012.1-1012.3.

Wang, H., Sun, S., Ge, W., Zhao, L., Hou, B., Wang, K., Lyu, Z., Chen, L., Xu, S., Guo, J., Li, M., Su, P., Li, X., Wang, G., Bo, C., Fang, X., Zhuang, W., Cheng, X., Wu, J. and Dong, L., 2020a. Horizontal gene transfer of Fhb from fungus underlies *Fusarium* head blight resistance in wheat. *Science*, 368(6493). doi:10.1126/science.aba5435.

Wang, J., Tian, L., Lee, H. and Chen, Z.J., 2006. Nonadditive regulation of *FRI* and *FLC* loci mediates flowering-time variation in *Arabidopsis* allopolyploids. *Genetics*, 173(2), pp.965-974. doi:10.1534/genetics.106.056580.

Wang, S., Guan, Y., Wang, Q., Zhao, J., Sun, G., Hu, X., Running, M.P., Sun, H. and Huang, J. (2020b). A mycorrhizae-like gene regulates stem cell and gametophore development in mosses. *Nature Communications*, 11(1), pp.1–11. doi:10.1038/s41467-020-15967-6.

Wang, Y., Bräutigam, A., Weber, A.P.M. and Zhu, X., 2014. Three distinct biochemical subtypes of C_4 photosynthesis? A modelling analysis. *Journal of Experimental Botany*, 65(13), pp.3567-3578.

Watcharamongkol, T., Christin, P.A. and Osborne, C.P., 2018. C_4 photosynthesis evolved in warm climates but promoted migration to cooler ones. *Ecology Letters*, 21(3), pp.376–383. doi:10.1111/ele.12905.

Wickell, D.A., and Li, F.W., 2019. On the evolutionary significance of horizontal gene transfers in plants. *New Phytologist*, 225(1), pp.113-117.

Williams, B.P., Aubry, S. and Hibberd, J.M., 2012. Molecular evolution of genes recruited into C_4 photosynthesis. *Trends in Plant Science*, 17(4), pp.213-220.

Williams, T.A., Cox, C.J., Foster, P.G., Szöllősi, G.J. and Embley, T.M. (2019). Phylogenomics provides robust support for a two-domains tree of life. *Nature Ecology & Evolution*, 4(1), pp.138–147. doi:10.1038/s41559-019-1040-x.

Worrell, V., Nagle, D., McCarthy, D. and Eisenbraun, A., 1988. Genetic transformation system in the archaebacterium *Methanobacterium thermoautotrophicum* Marburg. *Journal of Bacteriology,* 170(2), pp.653-656.

Wu, D., Jiang, B., Ye, C., Timko, M.P. and Fan, L., 2022. Horizontal transfer and evolution of the biosynthetic gene cluster for benzoxazinoids in plants. *Plant Communications*, 3(3), p.100320. doi:10.1016/j.xplc.2022.100320.

Xi, Z., Bradley, R.K., Wurdack, K.J., Wong, K.M., Sugumaran, M., Bomblies, K., Rest, J.S., and Davis, C.C., 2012. Horizontal transfer of expressed genes in a parasitic flowering plant. *BMC Genomics*, 13(1). pp.227. doi:10.1186/1471-2164-13-227.

Xia, J., Guo, Z., Yang, Z., Han, H., Wang, S., Xu, H., Yang, X., Yang, F., Wu, Q., Xie, W., Zhou, X., Dermauw, W., Turlings, T.C.J. and Zhang, Y., 2021. Whitefly hijacks a plant detoxification gene that neutralizes plant toxins. *Cell*, 184(7), pp.1693-1705.e17.

Yang, Z., Wafula, E.K., Kim, G., Shahid, S., McNeal, J.R., Ralph, P.E., Timilsena, P.R., Yu, W., Kelly, E.A., Zhang, H., Person, T.N., Altman, N.S., Axtell, M.J., Westwood, J.H. and dePamphilis, C.W. (2019). Convergent horizontal gene transfer and cross-talk of mobile nucleic acids in parasitic plants. *Nature Plants*, 5(9), pp.991–1001. doi:10.1038/s41477-019-0458-0.

Yanai, A., Wolf, Y.I., Koonin, E.V., 2002. Evolution of gene fusions: horizontal transfer versus independent events. *Genome Biology,* 3(5). Pp 0024.1 – 0024.13. doi: 10.1186/gb-2002-3-5-research0024

Yoshida S., Maruyama S., Nozaki H. and Shirasu K., 2010. Horizontal gene transfer by the parasitic plant *Striga hermonthica*. *Science* 328(5982), pp.1128-1128.

Yoshida, S., Kim, S., Wafula, E.K, Tanskanen, J., Kim, Y.M., Honaas, L., Yang, Z., Spallek, T., Conn, C., Ichihashi, Y., Cheong, K., Cui, S., Der, J., Gundlach, H., Jiao, Y., Hori, C., Ishida, J., Kasahara, H., Kiba, T., Kim, M., Koo, N., Laohavisit, A., Lee, Y., Lumba, S., McCourt, P., Mortimer, J., Mutuku, J., Nomura, T., Sasaki-Sekimoto, Y., Seto, Y., Wang, Y., Wakatake, T., Sakakibara, H., Demura, T., Yamaguchi, S., Yoneyama, K., Manabe, R., Nelson, D., Schulman, A., Timko, M., dePamphilis, C., Choi, D. and Shirasu, K., 2019. Genome sequence of *Striga asiatica* provides insight into the evolution of plant parasitism. *Current Biology,* 29(18), pp.3041-3052. doi:10.1016/j.cub.2019.07.086.

Yu, L., Boström, C., Franzenburg, S., Bayer, T., Dagan, T. and Reusch, T.B.H., 2020. Somatic genetic drift and multilevel selection in a clonal seagrass. *Nature Ecology & Evolution*, 4(7), pp.952-962.

Yue, J., Hu, X., Sun, H., Yang, Y. and Huang, J., 2012. Widespread impact of horizontal gene transfer on plant colonization of land. *Nature Communications,* 3(1). doi:10.1038/ncomms2148.

Zhang, J., Fu, X.X., Li, R.Q., Zhao, X., Liu, Y., Li, M.H., Zwaenepoel, A., Ma, H., Goffinet, B., Guan, Y.L., Xue, J.Y., Liao, Y.Y., Wang, Q.F., Wang, Q.H., Wang, J.Y., Zhang, G.Q., Wang, Z.W., Jia, Y., Wang, M.Z. and Dong, S.S., 2020. The hornwort genome and early land plant evolution. *Nature Plants*, 6(2), pp.107–118. doi:10.1038/s41477-019-0588-4.

Zhang, J., 2003. Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, 18(6), pp.292-298.

Zhang, J., 2006. Parallel adaptive origins of digestive RNases in Asian and African leaf monkeys. *Nature Genetics*, 38(7), pp.819-823.

Zhang, Z., Qu, C., Zhang, K., He, Y., Zhao, X., Yang, L., Zheng, Z., Ma, X., Wang, X., Wang, W., Wang, K., Li, D., Zhang, L., Zhang, X., Su, D., Chang, X., Zhou, M., Gao, D., Jiang, W. and Leliaert, F. (2020b). Adaptation to extreme Antarctic environments revealed by the genome of a sea ice green alga. *Current Biology,* 30(17), pp.3330-3341.e7. doi:10.1016/j.cub.2020.06.029.