Widespread lateral gene transfer among grasses

Samuel G. S. Hibdige[1], Pauline Raimondeau[1,2], Pascal-Antoine Christin[1], Luke T. Dunning[1*]

Affiliations:

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

[2] Present address: Laboratoire Évolution & Diversité Biologique (EDB UMR5174), CNRS, Institut de Recherche pour le Développement, F-31062 Toulouse, France

[*]**Corresponding author:** Luke T. Dunning; Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom; +44 (0) 1142220027; l.dunning@sheffield.ac.uk

Word count:

Introduction = 860 words

Materials and Methods = 1,572 words

Results = 1,542 words

Discussion = 1,682 words

Total = 5,656 words

Number of figures: 4

Colour figures: 2 (Figure 1 and Figure 2)

Number of Tables: 3

Supporting information: Yes

Summary

- Lateral gene transfer (LGT) has been documented in a broad range of prokaryotes and eukaryotes, and it can promote 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 in the group.

- 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 LGT 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 with overlapping distributions also occur, pointing to an additional role of biogeography.

- Overall, we show that LGT is a widespread phenomenon in grasses, which has moved functional genes across the entire grass family into domesticated and wild species alike. The dynamics of successful LGT appears to be dependent on both opportunity (co-occurrence and rhizomes) and compatibility (phylogenetic distance).

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

1. 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 being 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, *in press*), including plants (reviewed in: Richardson & Palmer, 2007; Gao *et al.*, 2014; Wickell & Li, 2019).

DNA has been transferred into plants from prokaryotes, fungi and viruses, with recipients in particular 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 LGT 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). 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 grass (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 they are the only group where multiple LGT recipients have been identified, 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). 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.*, *In press*). 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), the number of LGT is predicted to be higher in perennial and rhizomatous species. Conversely, if LGT happens through illegitimate pollination, the number of LGT may not vary with growth form or be higher in annuals that produce seeds more frequently. 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 determine whether functional characteristics or phylogenetic distance determine 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 clade (Oryzoideae and Pooideae) and three from the PACMAD clade (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 group. In addition, the species represent a variety of domestication statuses, life-history strategies, genome sizes, and ploidy levels (Table 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 LGT received by the Paniceae tribe, a group for which seven genomes are available, to (iii) test whether the number of LGT increases with phylogenetic relatedness. Our work represents the first systematic quantification of LGT among members of a large group of plants and sheds new light on the conditions that promote genetic exchanges across species boundaries in plants.

2. Materials and Methods

2.1 Detecting grass-to-grass LGT

We modified the approach previously used by Dunning *et al.*, (2019) to identify grass-to-grass LGT. Specifically, we did not use the initial mapping filtering step from Dunning *et al.* as it relied on the availability of high-coverage genome data for pairs of closely related species. In total, 17 genomes were scanned for LGT (Table 1), with all phylogenetic analyses based on coding sequences (total = 817,621 genes; mean per species 48,095 genes; SD = 26,764 genes).

In the first step, 37-taxa trees were constructed using data from the 17 grass genomes (Table 1), supplemented with transcriptome data for 20 additional species from across the grass family (Moreno-Villena *et al.*, 2018; Table S1). BLASTn was used to identify the top-match for each gene in the 36 other species with a minimum match length of 300bp (not necessarily a single continuous BLAST match). Nucleotide alignments were generated by aligning the BLASTn matching regions to the query sequence using the 'add fragments' parameter in MAAFT v7.427 (Katoh & 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; Table S2). For each alignment with ten species or more, a maximum-likelihood phylogenetic tree was inferred using PhyML v.20120412 (Guindon & Gascuel, 2003) with the GTR+G+I substitution model. Each topology was then mid-point rooted using the phytools package in R and perl scripts 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 mulitgene coalescent species tree analysis (Figure 1; Dunning *et al.*, 2019). The whole set of analyses were later repeated

to detect LGT between well supported subclades within the Paniceae, considering LGT received from two clades represented by two genomes and supported by most gene trees in previous analyses (i.e. Cenchrinae and Panicinae, Figure 1). To be considered as nested, the sister species of the query gene, and their combined sister group, 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 to verify that the nesting of the query sequence was supported by a minimum of 50% of the bootstrap replicates.

The second filtering step was applied to those candidates that passed the first phylogenetic filter. For those, 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 (Table S1). 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 MAAFT 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 less than three species within or 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. The tree inference was then repeated with 100 bootstraps, and the trees were again manually inspected, discarding candidates where the nesting was supported by <70% bootstrap replicates. Finally, BLASTx was

used to annotate the LGT candidates against the SwissProt database.

After these two successive filters, retained candidates were subjected to further validation. To verify the nesting of the candidate LGT 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 & Hasegawa, 2001) using the site-wise likelihood values generated by PhyML, and p-values were Bonferroni corrected to account for multiple testing. LGTs 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; 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.; Table S1) 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.

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 1); (ii) their nesting had to be well supported (≥ 70% bootstrap 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 for exemplar LGT). Phylogenetic trees and alignments are included as supplementary datasets. All analyses were preformed using publicly available data (Table S1).

2.2 Synteny analyses

Synteny analyses were performed for a subset of seven diploid grass genomes with high-quality reference genomes (*Brachypodium distachyon*, *Oryza sativa*, *Oropetium thomaeum*, *Panicum hallii*, *Setaria italica*, *Sorghum bicolor*, and *Zea mays*) using SynFind (Tang *et al.*, 2015). 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 (Table S3).

2.3 Analyses of replicate sequencing runs to check for potential contamination

Independently sequenced runs for each of the model species were screened for the presence of each LGT, as potential contaminations would not appear in multiple replicates. 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). 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).

2.4 Grass traits

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 (Bennett & Leitch, 2005), and climatic information from Watcharamongkol *et al.*, (2018). The climate data for *Oropetium thomaeum* was not included in Watcharamongkol *et al.*, (2018), and was therefore retrieved from GBIF (GBIF.org; 11th July 2019) GBIF Occurrence Download https://doi.org/10.15468/dl.wyhtoo) and WorldClim (Harris *et al.*, 2014; Fick & Hijmans, 2017) using the same methods. All statistical tests were preformed 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 1). The Kruskal-Wallis tests 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 LGT, we preformed phylogenetic generalized least squares (PGLS) to account for the relatedness between samples. The PGLS analysis was preformed 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 1). Individual and iterative models were performed, removing the least significant variable until only significant variables remained (p-value <0.05).

3. Results

3.1 LGT occurs in all lineages and functional types of grass

Out of the 817,612 grass genes from the 17 grass genomes (Table 1) screened, 55.89% had sufficient homologous grass sequences (≥ 10 taxa) for reliable phylogenetic reconstruction (Table 2 & Table S2), 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 grass genomes in this initial analysis (Table 2; full results Table S2). The number of LGT received varied among species (p-value < 0.01; Chi-square test; mean = 8.4; SD=9.0; range=0 − 34; Table S2), 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 LGT in *Alloteropsis semialata* (Dunning *et al.*, 2019) 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 species (Figure 1; Table 2). No LGT 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 3). We detected LGT in wild species, but also in major crops. For instance, maize (*Zea mays*) has 11 LGT received from Chloridoideae and Paniceae, while wheat (*Triticum aestivum*) has 10 LGT 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 (Table S2). Across all plant properties, LGT seems to be more abundant in perennial, rhizomatous and C_4 species (Figure 3). A phylogenetic generalized least squares (PGLS) analysis was conducted to test for an effect of all traits 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 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 R^2 = 0.243) and the number of genes tested (p-value = 0.038, adjusted R^2 = 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 R^2 = 0.652) identified three characteristics that jointly increased the number of LGT: 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). 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.

3.2 LGT 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). 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 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. We therefore focused on the identity of donors of LGT to Paniceae, the group with the highest number

of complete genomes.

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 LGT transferred from the Panicinae and Cenchrinae subgroups of Paniceae (each represented by two genomes) to other Paniceae, in addition to those received from other groups. In total, we identify 129 LGT across the seven Paniceae genomes, 35 of which were transferred from the Cenchrinae and Panicinae subgroups (Table 3; full results Table S4). 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 LGT given per species decreases with the phylogenetic distance to Paniceae, reaching lowest levels in the BOP clade (Pooideae and Oryzoideae; Figure 4).

3.3 Ruling out alternative hypotheses

There are four main alternative hypothesis to LGT: [1] incomplete lineage sorting, [2] unrecognised parology, [3] hybridisation, [4] contamination, and [5] phylogenetic bias, 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), 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 seven diploid species for this analysis, with at least one

representative from each of the five groups. For each LGT detected in these seven species, we determined whether the genes from the other six species identified as orthologous in the phylogenetic tree were syntenic to the LGT or the native gene. In no instance was a gene from another model species syntenic with the LGT, with 76.3% being syntenic with the native copy, and 23.7% being syntenic to neither (Table S3). The synteny analyses therefore 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). Indeed 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 13 genomes with LGT by confirming the presence of the laterally acquired DNA in multiple independent sequencing runs (Table S5 and Figure S1). 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. For the remaining 7 species, only the whole-genome data used to generate the reference assembly exists, with all but one (*Leersia perrieri*) having multiple sequencing libraries/runs. For each dataset, we compare the genome-wide mean coverage for each gene to that of the identified LGT (Table S5 and Figure S1). All LGTs had sequencing data in the multiple 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 (Table S5 and Figure S1). Overall, these results confirm that contamination in the original reference genomes is not responsible for the presence of the LGT 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 a high similarity between the putative donor and recipient on intron sequences of LGT in addition to exons (Figure 2). It was however not possible to delimit with high precision the laterally acquired fragments detected here (as done for *A. semialata* in 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. The observation of some intergenic regions with high similarity (Figure S2), together with intronic similarities (Figure 2), still 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.

4. Discussion

Lateral gene transfer is a potent evolutionary force capable of having a profound impact on the evolutionary trajectory of a species (Li et al., 2014; Cheng et al., 2019; Phansopa et al. In press). Here, we use grasses as a model system 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 LGT per 10,000 genes; 135 in the first round of analyses and 35 among groups of Paniceae) that have been laterally transferred to 13 of the 17 complete grass genomes that were screened (Table 1 & Table 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 LGT 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).

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 59 LGT were previously detected (Dunning *et al.*, 2019). Here, only 20 were identified when considering solely LGT among the five higher groups (Table 2), while a further 13 were detected when considering subgroups of Paniceae as potential donors (Table 3). The other LGTs previously identified for *A. semialata* were not detected in the present study because of slight methodological differences (e.g. less donor clades considered and sampling more genomes means greater multiple testing correction), and because we did not screen the flanking regions of LGTs for additional genes with high sequence similarity in the present study due to the larger scale of the present analyses,

varying genome assembly contiguity, and a lack of whole-genome data for some donor species. Conversely, improvements in the initial screening approach (i.e. removing an initial read-mapping step from Dunning *et. al*., 2019 method) resulted in the identification of four novel LGT. These examples prove that our ability to detect LGT depends on many factors and strongly suggest that the LGTs we report here concern only a small fraction of those existing in grass genomes. Finally, our approach precludes the detection of older LGT that are shared by multiple individuals within a clade. 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.

4.1 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 1). We detected LGT in domesticated and wild species alike (Figure 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 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. However, there is a statistical increase of the number of LGT in rhizomatous species and two of the three species with the highest numbers of LGT (*Alloteropsis semialata* and *Panicum virgatum*) are perennials that can propagate vegetatively via rhizomes (Table 1 & Table 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). In this instance, LGT is

similar to somatic mutations occurring in clonal species, as documented in the seagrass (*Zostera mariana*) 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 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.

4.2 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 LGT from closer relatives (Figure 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 inserted in non-syntenic positions in the genome where they coexist with the native copy, often on different chromosomes (Table S3). 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 are available (e.g. *Themeda triandra*; Dunning *et al.* 2019). 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 LGT from close relatives or to use it after the transfers, thereby increasing the chance of selectively retaining it.

4.3 The role of overlapping distributions.

In addition to genomic compatibility, the probability of co-occurrence in the wild might decrease with divergence time, explaining the observed effect of phylogenetic distance (Figure 4). Paniceae grasses generally grow in biodiverse savannas that are dominated by Andropogoneae species in wetter areas, and Chloridoideae grasses in drier habitats (Lehmann *et al.*, 2019). The effects of phylogenetic distance are therefore confounded with those of biogeography, but specific examples indicate that biogeography can take precedence.

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), 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* co-occurs 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 can not 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 LGT in *Panicum hallii* were

475 received from Chloridoideae, while a majority (81%) of those in *Panicum virgatum* were received
476 from Andropogoneae (Table 3). This pattern mirrors the dominant grassland type (Chloridoideae vs.
477 Andropogoneae) for a majority of the range of each of the two species, and the area from which the
478 individual for the genome assembly was sampled (Lovell *et al.*, 2018; Lehmann *et al.*, 2019).
479
480 Quantifying the effects of biogeography as opposed to other factors requires identifying the donor
481 to the species level and a detailed description of the spatial distribution of each grass species,
482 including their abundances. Indeed, the likelihood of encounters will increase with the number of
483 individuals of the donor species and not just its presence. Such ecological datasets coupled with
484 genomic data for a large numbers of grasses will be key to future studies of LGT dynamics in this
485 group of plants.
486
487 *4.4 Conclusions*
488 Using stringent phylogenomic filtering, we have shown that lateral gene transfer (LGT) is a
489 widespread process in grasses, where it occurs in wild species as well as in widely cultivated crops
490 (e.g. maize and wheat). LGT does not appear restricted to particular functional types, although it
491 seems to increase in rhizomatous species, where vegetative growth offers extra opportunities for
492 gene transfers into the germline. In addition, we show that the amount of successful transfers
493 decreases with phylogenetic distance. This effect of the phylogenetic distance might result from
494 increased genomic compatibility among more related groups. Alternatively, groups that diverged
495 more recently might be more likely to co-occur, offering more opportunities for genetic exchanges.
496 Indeed, biogeography seems to have an overall effect on the frequency of LGT, as the only species
497 that received genes from a group that diverged more than 50 million years ago is the one with an
498 overlapping distribution following a relatively recent niche shift. Overall, our study shows that LGT
499 occurs in a variety of grasses, and the frequent movements of functional genes can strongly impact

the evolution of this group of plants.

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).

6. Authors' contributions

All authors designed the project. LTD, SGSH and PR conducted the analyses. LTD, PAC and SGSH wrote the paper, with the help of PR.

7. References

Andersson JO. **2005.** Lateral gene transfer in eukaryotes. *Cellular and Molecular Life Sciences CMLS* **62**: 1182-1197.

Barrett RD, Schluter D. 2008. Adaptation from standing genetic variation. *Trends in Ecology & Evolution* **23** :38-44.

Bennett MD, Leitch IJ. 2005. Plant DNA C-values database. https://cvalues.science.kew.org/

Bennetzen JL, Schmutz J, Wang H, Percifield R, Hawkins J, Pontaroli AC, ... Jenkins J. 2012. Reference genome sequence of the model plant *Setaria*. *Nature Biotechnology* **30**: 555.

Bowman JL, Kohchi T, Yamato KT, Jenkins J, Shu S, Ishizaki K, Yamaoka S, Nishihama R, Nakamura Y, Berger F, Adam C. 2017. Insights into land plant evolution garnered from the *Marchantia polymorpha* genome. *Cell* **171**: 287-304.

Broad Institute. 2019. Picard Toolkit. GitHub Repository. http://broadinstitute.github.io/picard/

Cannarozzi G, Plaza-Wüthrich S, Esfeld K, Larti S, Wilson YS, Girma D, ... Lyons E. 2014. Genome and transcriptome sequencing identifies breeding targets in the orphan crop tef (*Eragrostis tef*). *BMC Genomics* **15**: 581.

Chang BS, Campbell DL. 2000. Bias in phylogenetic reconstruction of vertebrate rhodopsin sequences. *Molecular Biology and Evolution* **17**: 1220-31.

Chen F, Dong W, Zhang J, Guo X, Chen J, Wang Z, ... Zhang L. 2018. The sequenced angiosperm genomes and genome databases. *Frontiers in Plant Science* **9***:* 418.

Cheng S, Xian W, Fu Y, Marin B, Keller J, Wu T, Sun W, Li X, Xu Y, Zhang Y, Wittek S. 2019. Genomes of subaerial Zygnematophyceae provide insights into land plant evolution. *Cell* **179**: 1057-67.

Christin PA, Besnard G, Edwards EJ, Salamin N. 2012. Effect of genetic convergence on phylogenetic inference. *Molecular Phylogenetics and Evolution* **62**: 921-927.

Christin PA, Edwards EJ, Besnard G, Boxall SF, Gregory R, Kellogg EA, ... Osborne CP. 2012. Adaptive evolution of C_4 photosynthesis through recurrent lateral gene transfer. *Current Biology* **22**: 445-449.

Christin PA, Spriggs E, Osborne CP, Strömberg CA, Salamin N, Edwards EJ. 2014. Molecular dating, evolutionary rates, and the age of the grasses. *Systematic Biology* **63**: 153-165.

Clayton WD, Vorontsova MS, Harman KT, Williamson H. 2016. GrassBase - The Online World Grass Flora.

Doolittle WF. 1999. Lateral genomics. *Trends in Biochemical Sciences* **24**: M5-M8.

Dunning LT, Olofsson JK, Parisod C, Choudhury RR, Moreno-Villena JJ, Yang Y, ... Christin PA. 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**: 4416-4425.

Fick SE, Hijmans RJ. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology* ***37***: 4302-4315.

Freschi L, Vincent AT, Jeukens J, Emond-Rheault JG, Kukavica-Ibrulj I, Dupont MJ, ... Levesque RC. 2018. The *Pseudomonas aeruginosa* pan-genome provides new insights on its population structure, horizontal gene transfer, and pathogenicity. *Genome Biology and Evolution* **11**: 109-120.

Gao C, Ren X, Mason AS, Liu H, Xiao M, Li J, Fu D. 2014. Horizontal gene transfer in plants. *Functional & Integrative Genomics* **14**: 23-29.

Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, ... Hadley D. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296**: 92-100.

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

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology* **52**: 696-704.

Guo L, Qiu J, Ye C, Jin G, Mao L, Zhang H, ... Lin Z. 2017. *Echinochloa crus-galli* genome analysis provides insight into its adaptation and invasiveness as a weed. *Nature Communications* **8**: 1031.

Harris IPDJ, Jones PD, Osborn TJ, Lister DH. 2014. Updated high-resolution grids of monthly climatic observations–the CRU TS3. 10 Dataset. *International Journal of Climatology* **34**: 623-642.

Husnik F, McCutcheon JP. 2018. Functional horizontal gene transfer from bacteria to eukaryotes. *Nature Reviews Microbiology* **16**: 67.

International Barley Genome Sequencing Consortium. 2012. A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491**: 711.

International Brachypodium Initiative. 2010. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**: 763.

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

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution* **30**: 772-780.

Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. *Nature Reviews Genetics* **9**: 605.

Kellogg EA, Buell CR. 2009. Splendor in the grasses. *Plant Physiology* **149**: 1-3.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357-359.

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

Lehmann CE, Griffith DM, Simpson KJ, Anderson TM, Archibald S, Beerling DJ, ... Fox DL. 2019. Functional diversification enabled grassy biomes to fill global climate space. *BioRxiv* **583625**.

Li FW, Villarreal JC, Kelly S, Rothfels CJ, Melkonian M, Frangedakis E, ... Burge DO .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**: 6672-6677.

Lovell JT, Jenkins J, Lowry DB, Mamidi S, Sreedasyam A, Weng X, ... Gordon SP .2018. The genomic landscape of molecular responses to natural drought stress in *Panicum hallii*. *Nature Communications* **9**: 5213.

Mahelka V, Krak K, Kopecký D, Fehrer J, Šafář J, Bartoš J, Blattner FR. 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**: 1726-1731.

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

Moreno-Villena JJ, Dunning LT, Osborne CP, Christin PA. 2017. Highly expressed genes are preferentially co-opted for C_4 photosynthesis. *Molecular Biology and Evolution* **35**: 94-106.

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

643 **Olofsson JK, Dunning LT, Lundgren MR, Barton HJ, Thompson J, Cuff N, Ariyarathne M,**
644 **Yakandawala D, Sotelo G, Zeng K, Osborne CP. 2019.** Population-specific selection on
645 standing variation generated by lateral gene transfers in a grass. *Current Biology* **29**: 3921-7.

646 **Orme D, Freckleton R, Thomas G, Petzoldt T, Fritz S. 2013.** The caper package: comparative
647 analysis of phylogenetics and evolution in R. *R package version 5:* 1-36.

648 **Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, ... Schmutz J.**
649 **2009.** The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**: 551.

650 **Peccoud J, Loiseau V, Cordaux R, & Gilbert C. 2017.** Massive horizontal transfer of transposable
651 elements in insects. *Proceedings of the National Academy of Sciences of the United States of*
652 *America* **114**: 4721-4726.

653 **Phansopa C, Dunning LT, Reid JD, Christin PA. In Press.** Lateral gene transfer acts as an
654 evolutionary shortcut to efficient C4 biochemistry. *Molecular Biology and Evolution*
655 doi:10/1093/molbev/msaa143

656 **Popa O & Dagan T. 2011.** Trends and barriers to lateral gene transfer in prokaryotes. *Current*
657 *Opinion in Microbiology* **14**: 615-623.

658 **Prentice HC, Li Y, Lönn M, Tunlid A, Ghatnekar L 2015.** A horizontally transferred nuclear gene
659 is associated with microhabitat variation in a natural plant population. P*roceedings of the*
660 *Royal Society B: Biological Sciences* **282**: 20152453.

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

663 **Richardson AO, Palmer JD. 2007.** Horizontal gene transfer in plants. *Journal of Experimental*
664 *Botany* **58**: 1-9.

665 **Schönknecht G, Weber AP, Lercher MJ. 2014.** Horizontal gene acquisitions by eukaryotes as
666 drivers of adaptive evolution. *Bioessays* **36**: 9-20.

667 **Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, ... Minx P. 2009.** The B73
668 maize genome: complexity, diversity, and dynamics. *Science* **326**: 1112-1115.

669 **Soreng RJ, Peterson PM, Romaschenko K, Davidse G, Zuloaga FO, Judziewicz EJ, ...**
670 **Morrone O. 2015.** A worldwide phylogenetic classification of the Poaceae (Gramineae).
671 *Journal of Systematics and Evolution* **53**: 117-137.

672 **Soucy SM, Huang J, Gogarten JP. 2015.** Horizontal gene transfer: building the web of life.
673 *Nature Reviews Genetics* **16**: 472.

674 **Shimodaira H, Hasegawa M. 2001.** CONSEL: for assessing the confidence of phylogenetic tree
675 selection. *Bioinformatics* **17**: 1246-1247.

676 **Skippington E, Ragan MA. 2012.** Phylogeny rather than ecology or lifestyle biases the
677 construction of *Escherichia coli–Shigella* genetic exchange communities. *Open Biology* **2**:
678 120112.

679 **Stamatakis A. 2014.** RAxML version 8: a tool for phylogenetic analysis and post-analysis of large
680 phylogenies. *Bioinformatics* **30**: 1312-1313.

681 **Stegemann S, Bock R. 2009.** Exchange of genetic material between cells in plant tissue grafts.
682 *Science* **324**: 649-51.

683 S**tein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang C, ... Wei S. 2018.** Genomes of 13
684 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation
685 across the genus *Oryza*. *Nature Genetics* **50**: 285.

686 **Studer AJ, Schnable JC, Weissmann S, Kolbe AR, McKain MR, Shao Y, ... Brutnell TP. 2016.**
687 The draft genome of the C_3 panicoid grass species *Dichanthelium oligosanthes*. *Genome*
688 *Biology* **17**: 223.

Tanaka H, Hirakawa H, Kosugi S, Nakayama S, Ono A, Watanabe A, ... Shimizu K. 2016. Sequencing and comparative analyses of the genomes of zoysiagrasses. *DNA Research* **23**: 171-180.

Tang H, Bomhoff MD, Briones E, Zhang L, Schnable JC, Lyons E. 2015. SynFind: compiling syntenic regions across any set of genomes on demand. *Genome Biology and Evolution* **7**: 3286-3298.

Vallenback P, Jaarola M, Ghatnekar L, Bengtsson BO. 2008. Origin and timing of the horizontal transfer of a *PgiC* gene from *Poa* to *Festuca ovina*. *Molecular Phylogenetics and Evolution* **46**: 890-896.

Van Etten J, Bhattacharya D. In press. Horizontal Gene Transfer in Eukaryotes: Not if, but How Much? *Trends in Genetics* doi:10.1016/j.tig.2020.08.006

VanBuren R, Bryant D, Edger PP, Tang H, Burgess D, Challabathula D, Freeling M. 2015. Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature* **527**: 508.

Varshney RK, Shi C, Thudi M, Mariac C, Wallace J, Qi P, ... Srivastava RK. 2017. Pearl millet genome sequence provides a resource to improve agronomic traits in arid environments. *Nature Biotechnology* **35**: 969.

Vogel A, Schwacke R, Denton AK, Usadel B, Hollmann J, Fischer K, ... Mayer KF. 2018. Footprints of parasitism in the genome of the parasitic flowering plant *Cuscuta campestris*. *Nature Communications* **9**: 2515.

Watcharamongkol T, Christin PA, Osborne CP. 2018. C_4 photosynthesis evolved in warm climates but promoted migration to cooler ones. *Ecology Letters* **21**: 376-383.

Wickell DA, Li F-W. 2019. On the evolutionary significance of horizontal gene transfer in plants. *New Phytologist* **225**: 113-117.

Xi Z, Bradley RK, Wurdack KJ, Wong KM, Sugumaran M, Bomblies K, ... Davis CC. 2012. Horizontal transfer of expressed genes in a parasitic flowering plant. *BMC Genomics* **13**: 227.

Yang Z, Wafula EK, Kim G, Shahid S, McNeal JR, Ralph PE, ... Person TN. 2019. Convergent horizontal gene transfer and cross-talk of mobile nucleic acids in parasitic plants. *Nature Plants* **5**: 991-1001.

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

Yu L, Boström C, Franzenburg S, Bayer T, Dagan T, Reusch TB. 2020. Somatic genetic drift and multilevel selection in a clonal seagrass. *Nature Ecology & Evolution* 1-11.

Yue J, Hu X, Sun H, Yang Y, Huang J. 2012. Widespread impact of horizontal gene transfer on plant colonization of land. *Nature Communications* **3**: 1-9.

Zhang J, Fu XX, Li RQ, Zhao X, Liu Y, Li MH, Zwaenepoel A, Ma H, Goffinet B, Guan YL, Xue JY. 2020. The hornwort genome and early land plant evolution. *Nature plants* **6**:107-18.

735 **Tables**

736 **Table 1: Model species used in this study and associated traits.**

Clade	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

737

738 Ploid. = Ploidy; 1C = 1C genome size in Gb; Cult. = cultivated (Y = yes; N = no); LH = life history
739 (A = annual; P = perennial); Clim. = climate (Temp = temperate; Trop = tropical); Phot. =
740 photosynthetic type; Cont. = number of continents; Rhiz. = rhizomatous (Y = yes; N = no).
741 [1]International Brachypodium Initiative, 2010; [2]International Barley Genome Sequencing
742 Consortium, 2012; [3]International Wheat Genome Sequencing Consortium, 2014; [4]Goff *et al.*,
743 2002; [5]Stein *et al.*, 2018; [6]Cannarozzi *et al.*, 2014; [7]VanBuren *et al.*, 2015; [8]Tanaka *et al.*, 2016;
744 [9]Patterson *et al.*, 2009; [10]Schnable *et al.*, 2009; [11]Dunning *et al.*, 2019; [12]Varshney *et al.*, 2017;
745 [3]Studer *et al.*, 2016; [14]Guo *et al.*, 2017; [15]Lovell *et al.*, 2018; [16]*Panicum virgatum* v4.1, DOE-JGI,
746 http://phytozome.jgi.doe.gov/; [17]Bennetzen *et al.*, 2012

747

748 **Table 2: Number of lateral gene transfers (LGT) detected between the five clades.**

			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	-

749

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

754

755

756

757

758

759

760

761 **Table 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

762

763 The number of species in each clade is indicated in parentheses, with values from Soreng et al.,
764 (2015); Pooid. = Pooideae; Ory. = Oryzoideae; Chlor. = Chloridoideae; Andro. = Andropogoneae;
765 Cench. = Cenchrinae; Pani. = Panicinae.

766

767

768

769

770

771

772

773

774

775

776

777

778

779

Figures

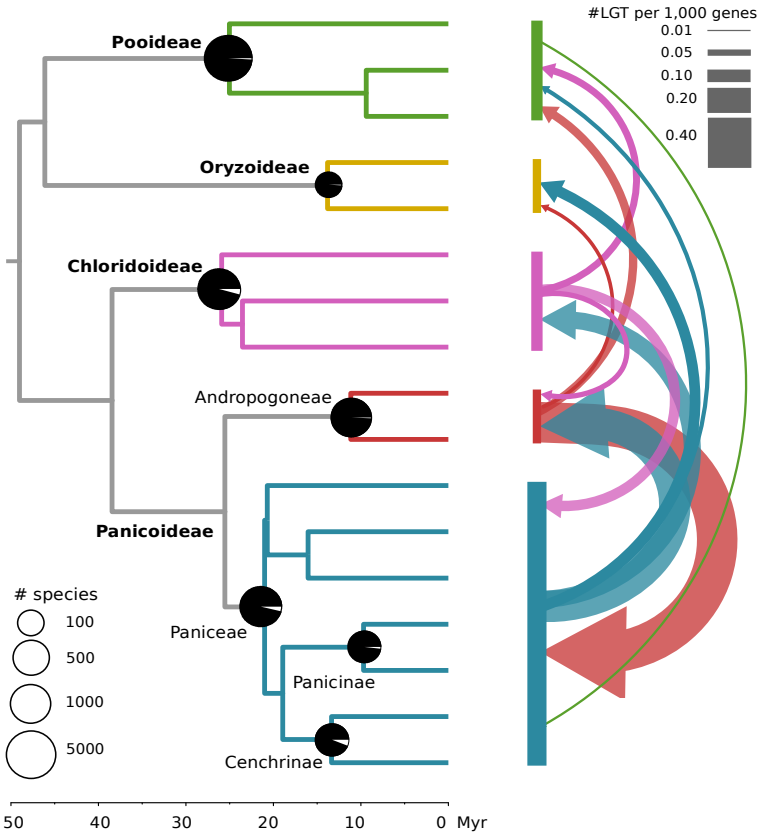

Fig 1: Distribution of lateral gene transfers among grasses.

Time-calibrated phylogenetic tree of 17 model grass species used in this study (extracted 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 LGT 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 LGT detected in each genome.

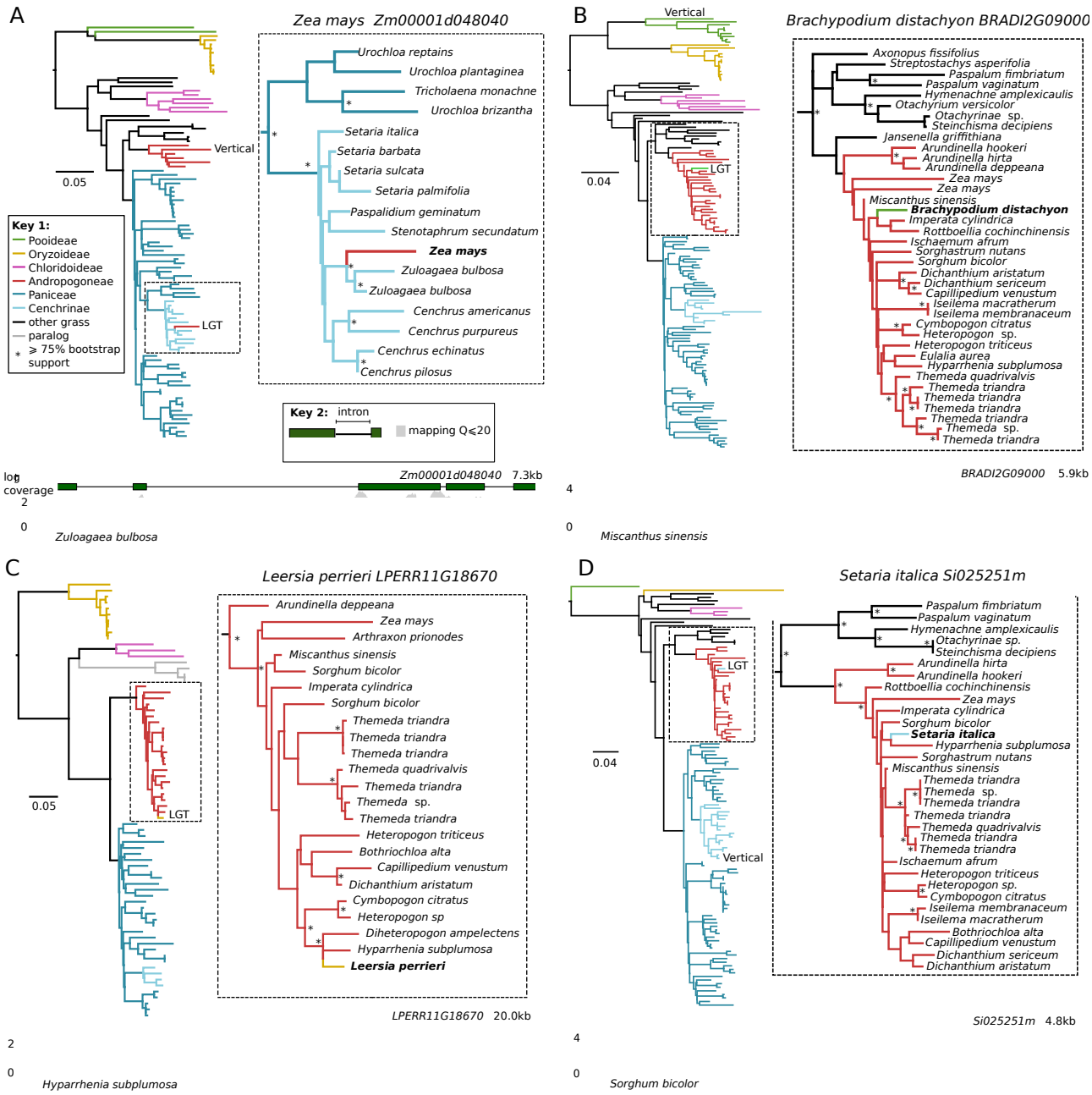

Fig 2: Four examples of grass-to-grass lateral gene transfer.

Each panel (A-D) shows an exemplar grass-to-grass LGT, with full and expanded regions of maximum likelihood phylogenies shown. 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.

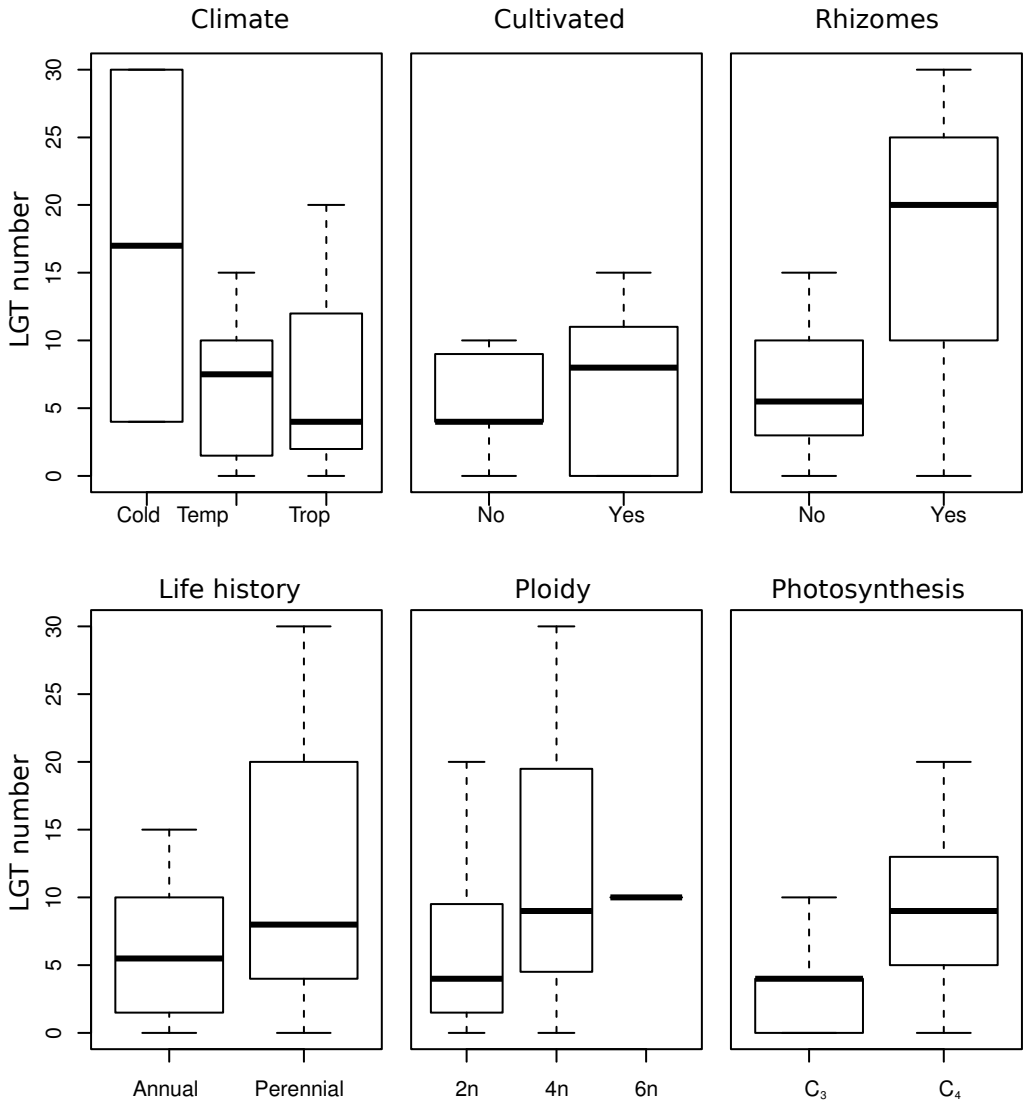

Fig. 3: Numbers of lateral gene transfer (LGT) received by different categories of grasses. For each group, the distribution of LGT numbers is shown with box plots connecting the median and the interquartile range, with whiskers showing 1.5 x the interquartile range.

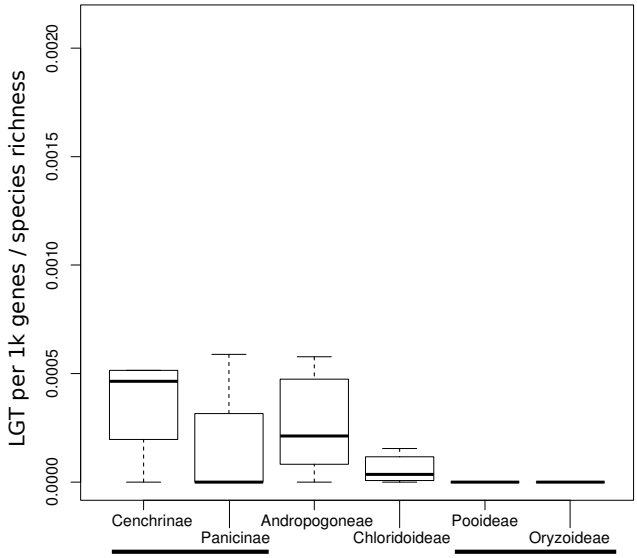

Fig. 4: Number of lateral gene transfer (LGT) received by Paniceae species from different clades.

The number of LGT 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.