White Rose Research Online

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk
Universities of Leeds, Sheffield and York

Deposited via The University of Sheffield.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/id/eprint/204080/

Version: Published Version

Article:
Raimondeau, P., Bianconi, M.E., Pereira, L. et al. (2023) Lateral gene transfer generates accessory genes that accumulate at different rates within a grass lineage. New Phytologist, 240 (5). pp. 2072-2084. ISSN: 0028-646X

https://doi.org/10.1111/nph.19272

Reuse
This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:
https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

UNIVERSITY OF LEEDS
University of Sheffield
UNIVERSITY of York

New Phytologist

Research

Lateral gene transfer generates accessory genes that accumulate at different rates within a grass lineage

Pauline Raimondeau[1,2], Matheus E. Bianconi[1], Lara Pereira[1], Christian Parisod[3], Pascal-Antoine Christin[1,3] and Luke T. Dunning[1]

[1]Ecology and Evolutionary Biology, School of Biosciences, University of Sheffield, Western Bank, Sheffield, S10 2TN, UK; [2]Laboratoire Evolution et Diversité Biologique, UMR5174, CNRS/IRD/Université Toulouse 3, Toulouse, 31062, France; [3]Department of Biology, University of Fribourg, Chemin du Musée 10, Fribourg, 1700, Switzerland

Author for correspondence:
Luke T. Dunning
Email: *l.dunning@sheffield.ac.uk*

Received: *5 June 2023*
Accepted: *30 August 2023*

New Phytologist (2023)
doi: 10.1111/nph.19272

Key words: adaptation, evolution, horizontal gene transfer, phylogenomics, Poaceae.

Summary

- Lateral gene transfer (LGT) is the movement of DNA between organisms without sexual reproduction. The acquired genes represent genetic novelties that have independently evolved in the donor's genome. Phylogenetic methods have shown that LGT is widespread across the entire grass family, although we know little about the underlying dynamics.
- We identify laterally acquired genes in five *de novo* reference genomes from the same grass genus (four *Alloteropsis semialata* and one *Alloteropsis angusta*). Using additional resequencing data for a further 40 *Alloteropsis* individuals, we place the acquisition of each gene onto a phylogeny using stochastic character mapping, and then infer rates of gains and losses.
- We detect 168 laterally acquired genes in the five reference genomes (32–100 per genome). Exponential decay models indicate that the rate of LGT acquisitions (6–28 per Ma) and subsequent losses (11–24% per Ma) varied significantly among lineages. Laterally acquired genes were lost at a higher rate than vertically inherited loci (0.02–0.8% per Ma).
- This high turnover creates intraspecific gene content variation, with a preponderance of them occurring as accessory genes in the *Alloteropsis* pangenome. This rapid turnover generates standing variation that can ultimately fuel local adaptation.

Introduction

Genomes are dynamic, with continuous gene loss compensated by gene duplication, and occasional *de novo* gene formation (Puigbò *et al.*, 2014; Schlötterer, 2015; Murat *et al.*, 2017; Fernández & Gabaldón, 2020). Because these processes happen at the individual level, they lead to intraspecific variation in gene content. In prokaryotes, this variation in gene content between individuals is dramatic, and its discovery led to the concept of the pangenome (Tettelin *et al.*, 2005). In a pangenome, genes are either defined as core, being present in all individuals of a species, or accessory and only present in a subset of individuals. One of the main driving forces for free-living prokaryote pangenome evolution is lateral gene transfer (LGT). LGT is the acquisition of DNA without sexual reproduction and this process continually adds genetic novelty to a species gene pool (Puigbò *et al.*, 2014; Brockhurst *et al.*, 2019). Pangenomes were more recently established for several groups of eukaryotes, where significant gene content variation was also observed with important ramifications for adaptation (Gordon *et al.*, 2017; Golicz *et al.*, 2020; Tao *et al.*, 2021).

The occurrence of LGT in eukaryotes is now widely accepted (Van Etten & Bhattacharya, 2020), and its relative contribution to the pangenome has been studied in a few groups containing mainly unicellular organisms. Less than 0.5% of the genes in the pangenome of yeast (*Saccharomyces cerevisiae*) have been acquired through LGT (Soanes & Richards, 2014), with many of these genes appearing as accessory loci (Han *et al.*, 2021). Similar proportions have been observed in comparative genomic studies across phytoplankton and within other algal groups where LGT accounts for 0.16–1.44% of genes in the genome (Fan *et al.*, 2020). The distribution of laterally acquired genes in these groups also revealed continuous gene transfers during their diversification (Dorrell *et al.*, 2021). LGT also happens in some multicellular eukaryotes (Keeling & Palmer, 2008), with unequivocal examples in fungi (Reynolds *et al.*, 2018), animals (Xia *et al.*, 2021) and plants (El Baidouri *et al.*, 2014; Li *et al.*, 2014, 2018; Wang *et al.*, 2020; Wickell & Li, 2020; Ma *et al.*, 2022). In terms of plant-to-plant transfers, LGT is especially prevalent between parasites and their hosts (Yoshida *et al.*, 2010; Kado & Innan, 2018; Cai *et al.*, 2021), and in some nonparasitic groups, such as grasses (Poaceae; Mahelka *et al.*, 2017; Dunning *et al.*, 2019; Hibdige *et al.*, 2021; Mahelka *et al.*, 2021; Wu *et al.*, 2022). Some of the laterally acquired genes received by plants have had drastic impacts on environmental adaptation (Li *et al.*, 2014; Phansopa *et al.*, 2020; Wang *et al.*, 2020), but their temporal dynamics and contribution to the pangenome remain understudied.

Among grasses, *Alloteropsis semialata* (Paniceae tribe of Panicoideae subfamily) represents one of the best study systems to investigate grass-to-grass LGT as it has the highest number of laterally acquired genes known for a diploid grass (Hibdige

© 2023 The Authors
New Phytologist © 2023 New Phytologist Foundation
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

et al., 2021). Up to 59 laterally acquired genes were present in the genome of a single Australian *A. semialata* accession, after ruling out alternative hypotheses such as contamination, incomplete lineages sorting, unrecognised paralogy, phylogenetic bias and hybridisation (Table 1). These genes were acquired from at least nine different donors separated by 20–40 Myr of evolution and multiple speciation events that gave rise to thousands of descendant species (Dunning *et al.*, 2019; Hibdige *et al.*, 2021). *Alloteropsis semialata* originated in tropical Africa, where divergent genetic lineages and the sister species *Alloteropsis angusta* still occur (Bianconi *et al.*, 2020). Previous analyses have identified laterally acquired genes that are present in multiple *Alloteropsis* species, that were either acquired before the divergence of the species or that were subsequently introgressed after their speciation (Olofsson *et al.*, 2016, 2019; Dunning *et al.*, 2019). However, the reliance on a single reference genome constrained previous systematic detection efforts to laterally acquired genes present in this sequenced individual. Quantifying the rate of LGT and its contribution to the gene content of a species requires considering multiple reference genomes for a diverse set of accessions, estimating the ages of the transfer based on their distribution among accessions and identifying any subsequent loss of the laterally transferred genes.

In this study, we generate complete reference genomes for three accessions of *A. semialata* representing various African sublineages and one for *A. angusta*, leading to five reference genomes for diploid individuals with the inclusion of the original Australian reference (Dunning *et al.*, 2019). We use phylogenetic approaches to identify all protein-coding genes laterally acquired from other grasses in each of the five genomes. We then use whole-genome sequence data for 40 additional diploid *Alloteropsis* accessions to (1) establish the distribution of all identified laterally acquired genes across the diversity of the group and map their origins onto a time-calibrated phylogeny, testing for the contributions of shared history and close geographical proximity on the sorting of these genes. The timing of the acquisition of each laterally acquired gene is then fitted to an exponential decay model to (2) directly estimate the rate of LGT gains and subsequent losses of the laterally acquired genes in the line of ancestors leading to each of the five reference genomes, and to contrast the latter with the rate of losses of vertically inherited genes. Finally, (3) we compare the intraspecific variation in laterally acquired gene content to the amount of variation of native genes to quantify the contribution of LGT to the pangenome of the group.

Materials and Methods

Genome sequencing, assembly and annotation

This study uses five *de novo* assembled *Alloteropsis* genomes, four of which were sequenced as part of this study. This includes three *A. semialata* (R. Br.) Hitchc. assemblies that, together with the previously sequenced genome from an Australian individual (accession AUS1; Clade IV; Dunning *et al.*, 2019), encompass the four main nuclear clades within this species (Fig. 1; Bianconi *et al.*, 2020): one individual from South Africa (accession RSA5-3; Clade I), one from Tanzania (accession TAN1-04B; Clade II) and one from Zambia (accession ZAM1505-10; Clade III). We also generated an assembly for *A. angusta* Stapf from a Ugandan accession (AANG_UGA4).

DNA was extracted from live plants grown at The University of Sheffield using the DNeasy Maxi Kit (Qiagen). Short-read Illumina library preparation and sequencing (HiSeq 2500 or 3000) was undertaken at the Edinburgh Genomics Centre (see Supporting Information Table S1 for per sample sequencing details). Long-read PacBio sequencing was generated using the SMRT PacBio platform at the Centre for Genomic Research at the University of Liverpool (see Table S1 for per sample sequencing details). Raw sequencing data were cleaned, assembled and annotated using the same approach as Dunning *et al.* (2019), full details are provided in Methods S1.

Identification of laterally acquired genes

For each of the five reference genomes, we independently detected all protein-coding laterally acquired genes using an

Table 1 Ruling out alternative hypotheses to lateral gene transfer.

There are five main alternative explanations and we cite evidence that largely rules these out. This evidence comes from numerous grasses, including the Australian *Alloteropsis semialata* used in this present study
(1) Contamination: The laterally acquired genes are present in multiple independent sequencing runs, including those generated by independent labs (Hibdige *et al*., 2021), and long-read sequencing confirms the LGTs are integrated into the nuclear genome of *A. semialata* (Dunning *et al*., 2019)
(2) Incomplete lineage sorting: A majority of known grass-to-grass LGTs (79.4%) are inserted in addition to the native ortholog (Hibdige *et al*., 2021). The appearance of both in the same gene tree precludes incomplete lineage sorting driving the observed patterns (as long as they are true orthologs, to be described later)
(3) Unrecognised parology: By comparing the synteny of orthologs from multiple model species across the gene trees it was shown that 76.2% of model grass orthologues were syntenic with the LGT recipients native copy (vertically inherited copy with an evolutionary history that tracks the species tree), 2.86% were syntenic to the laterally acquired gene, and 20.9% were syntenic to neither (Hibdige *et al*., 2021). The 2.86% could be a result of misassembly or homologous replacement (Dunning *et al*., 2019; Hibdige *et al*., 2021). The synteny analyses confirm that our method identifies true orthologues in most cases
(4) Phylogenetic bias (including convergent evolution): In addition to confirming the phylogenetic patterns with trees constructed on different data partitions (Dunning *et al*., 2019; Hibdige *et al*., 2021) there is also extremely high similarity of noncoding intergenic DNA between donor and recipient, something that would not be expected if the tree discordance was driven by convergent evolution or other phylogenetic biases. For example, a 45.7-kb noncoding region in the Australian *A. semialata* reference genome is 97.2% identical with that of the putative donor species (Dunning *et al*., 2019)
(5) Hybridisation: The coexistence and lack of synteny between the native and laterally acquired genes in the recipient genome argues against hybridisation through sexual reproduction and chromosomal recombination during the transfers (Dunning *et al*., 2019; Hibdige *et al*., 2021)

© 2023 The Authors
New Phytologist © 2023 New Phytologist Foundation

14698137, 0, Downloaded from https://nph.onlinelibrary.wiley.com/doi/10.1111/nph.19272 by Test, Wiley Online Library on [09/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

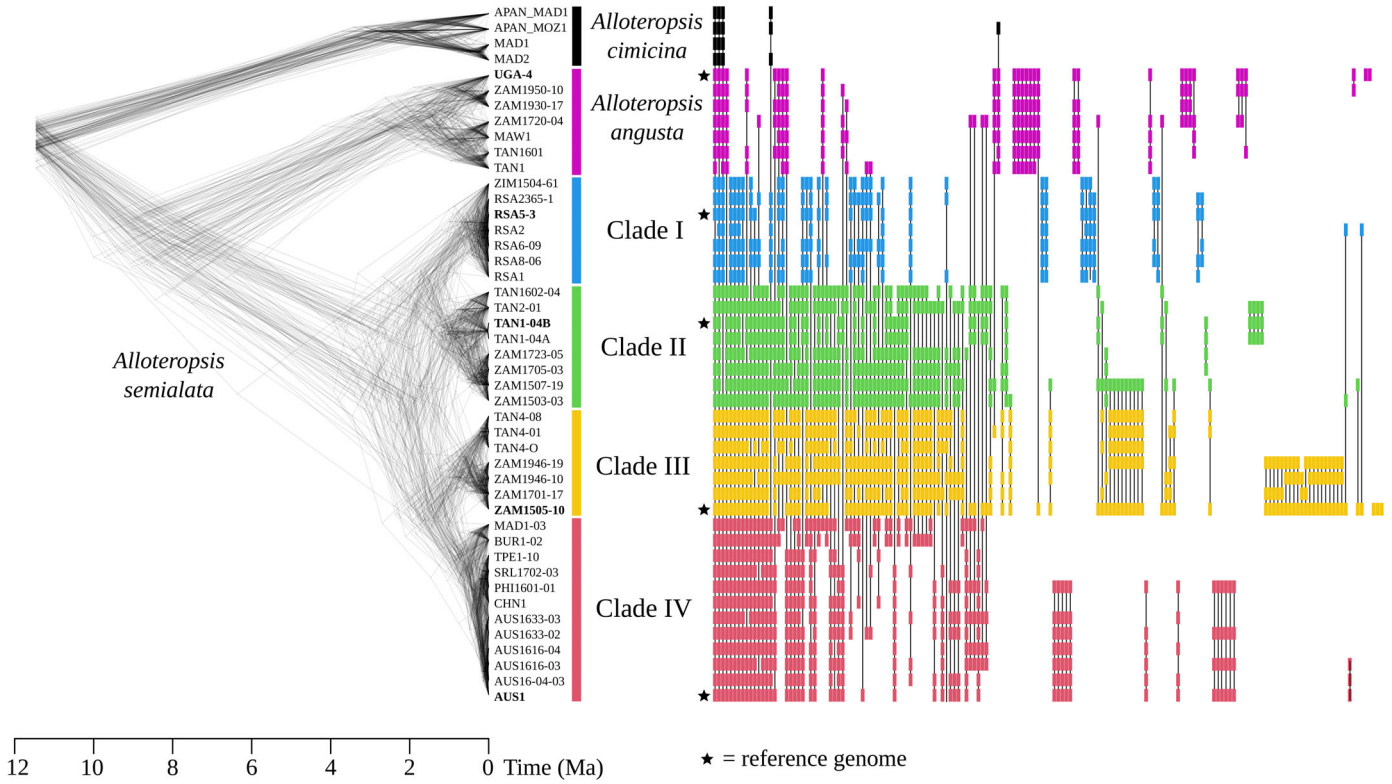

Fig. 1 Phylogenetic distribution of laterally acquired genes in *Alloteropsis*. Phylogenetic relationships of the sample used in this study are shown on the left, with each time-calibrated phylogenetic tree based on a different set of five nuclear genes. The sample names include a three-letter code denoting their country of origin, reference genome individuals are in bold, and the species and nuclear clades within *Alloteropsis semialata* are indicated. On the right, the presence of each laterally acquired gene in each accession is shown with a rectangle coloured according to the clade the sample belongs to. Each individual laterally acquired gene is connected by a vertical line and they are ordered by their abundance across all groups.

approach we previously developed (Dunning *et al.*, 2019; Hibdige *et al.*, 2021), which identifies *Alloteropsis* genes nested in distantly related clades of grasses that are resolved in a coalescence species tree analysis (Dunning *et al.*, 2019). In brief, identifying primary laterally acquired genes involves building phylogenetic trees of increasing species representation (up to 135 species; Table S2) and applying successive filters to verify that a scenario of LGT is statistically supported (Fig. S1). Secondary candidates are then identified as genes in close physical proximity to the primary candidates in the genome, and that have gene tree topologies supporting the same LGT scenario. The identification of secondary laterally acquired genes effectively rescues loci discarded by the stringent primary scan. All subsequent analyses made no distinction between primary and secondary laterally acquired genes. Full details of the method are provided in the Methods S1.

As the method was applied to each genome independently, older laterally acquired genes shared among accessions would potentially be detected in multiple scans. We therefore generated a merged list ensuring the occurrence of each gene was only counted once. Our method also counts recent duplicates (i.e. defined as being part of the same monophyletic group in the phylogeny) as a single LGT. This negates the problem of having to determine whether the gene duplicate arose prior or post lateral transfer, or indeed whether the duplication is in fact an assembly error (e.g. uncollapsed haplotypes). The function of the laterally acquired genes was investigated using a Gene Ontology (GO) enrichment analysis (see Methods S1 for details).

Distribution of laterally acquired genes across *Alloteropsis*

Genome assemblies and annotations are seldom complete, and there will often be missing loci. To reduce the effect of assembly and annotation quality on our results, we decided to verify whether each laterally acquired gene was present in the unassembled sequencing reads using a phylogenetic approach. From the five genomes, we selected a single loci to act as a representative for each laterally acquired gene ($n = 177$). The representative was selected based on the original gene tree alignments, with a preference for genes that yielded alignments with the most *Alloteropsis* sequences present, had the most taxa and were the longest. In all but two cases, a single reference gene was sufficient to recover all loci in the laterally acquired clade, with two nonoverlapping annotated sequences used as references where this was not possible. We then determined the presence of each of these representative genes in whole-genome short-read datasets belonging to 45 diploid *Alloteropsis* accessions (including the five individuals with reference genomes; Bianconi *et al.*, 2020), using a combination of BLASTN searches (minimum alignment length 100 bp) and phylogenetic analyses. For each of the 45 datasets, putative reads corresponding to each laterally acquired gene were identified via a BLASTN analysis with default parameters. The 10

© 2023 The Authors
New Phytologist © 2023 New Phytologist Foundation

14698137, 0 Downloaded from https://nph.onlinelibrary.wiley.com/doi/10.1111/nph.19272 by Test, Wiley Online Library on [09/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

top hits were retrieved, and each one was successively added to the gene tree alignments using MAFFT v.7.427 (Katoh & Standley, 2013) with the 'add fragments' parameter. A phylogenetic tree was then inferred with PHYML, with the best substitution model identified using Smart Model Selection SMS v.1.8.1 (Lefort *et al.*, 2017). Reads were considered as belonging to the laterally acquired gene if they were sister to it in this phylogenetic tree. Finally, the laterally acquired gene was considered as present if it was supported by at least three such reads. For the two genes with two nonoverlapping references, it was considered as present if either of the reference fragments fulfilled the criteria.

Molecular dating of *Alloteropsis*

We generated 100 different dated species phylogenetic trees that could be used to retrace the evolutionary history of each laterally acquired gene. Each tree was inferred from five randomly selected Benchmarking Universal Single-Copy Orthologs (BUSCO) gene alignments and time-calibrated species phylogenetic trees were inferred under a coalescence model, using Bayesian inference as implemented in *BEAST2 v.2.6.4 (Bouckaert *et al.*, 2019). Full details of the molecular dating methods are provided in Methods S1.

Testing for an effect of history and geography on the distribution of laterally acquired genes

For each pair of individuals, a similarity index was computed as the number of shared laterally acquired genes present in their genome. These similarities were first compared with their pairwise divergence times, using a Mantel test. The residuals of this relationship, which represent the part of the similarity index not explained by divergence times, were then tested for an effect of pairwise geographical distance, using the partial Mantel test. For each pair of individuals, the most recent divergence across the 100 phylogenetic trees was considered. The pairwise geographic distance along the Earth's surface was computed from GPS coordinates, using the earth.dist function in the R package fossil (Vavrek, 2011). For the Mantel test, the observed Spearman correlation coefficient was compared with those obtained with 9999 permutated matrices. The R^2 was extracted from a linear model.

Inferring the rates that laterally acquired genes are gained and lost

To estimate the origin, each laterally acquired gene was recoded as a presence/absence character and mapped onto a time-calibrated phylogenetic tree. These analyses were performed on a per gene basis and not at the genomic block level because genes can be independently lost and blocks in less contiguous genomes will be more fragmented. To allow each gene to evolve along the phylogenetic tree that best explained its history, the likelihood was estimated using the ace function with an asymmetrical substitution matrix (ARD model) in the APE package (Paradis & Schliep, 2019) for all 100 phylogenetic trees, and the tree producing the highest likelihood was selected. Stochastic mapping was then performed to map the origin of each laterally acquired gene using the make.simmap function in the PHYTOOLS package (Revell, 2012). If more than one acquisition was inferred along these branches, the history of gene was considered as ambiguous and it not included in the rate analyses. For genes where a single origin was identified, we extracted the date of acquisition. This analysis was performed independently for each of the five reference genomes, and the process was repeated 100 times, producing 100 sets of time of origin for each of the laterally acquired genes. These repeats represent pseudoreplicates. Note that, for a given gene, the number of acquisitions can vary among repeats of the stochastic mapping, so that it is considered in some but not all repeats.

The number of LGT per million-year time slices was extracted from the distribution of times of origin, up to 11 Ma, which represents the time to most recent common ancestor of the species of *Alloteropsis* included here. For each time slice, the number of LGTs averaged across the 100 replicates was used to estimate the rates of gains (G) and subsequent losses (L). For this purpose, an exponential decay equation was fitted using the nls function in R:

$$O = G \times (1-L)^t \qquad \text{Eqn 1}$$

O is the number of genes laterally acquired in each time slice that were retained until the present and detected in the reference genome, and t is the average age of the time window. The rates G and L were estimated using the same approach independently for each of the 100 pseudoreplicates of stochastic mapping, and the 0.025 and 0.975 quantiles of their distributions were used to compute 95% confidence intervals.

Analyses of native genes

Native genes are those that have been vertically inherited following the species tree and that do not have a LGT in their history (at least since the origin of the grass family). From the initial 37-taxa trees, we extracted gene trees that were unlikely to have undergone LGT in their recent history. To be considered as such, a gene had to be: present in at least the more distantly related congeneric *Alloteropsis Cimicina* (L.) Stapf and one of *A. angusta* and *A. semialata*; the *Alloteropsis* sequences had to be monophyletic; and they had to be placed in the phylogenetic tree within Paniceae (the tribe that contains *Alloteropsis*), but outside of the subtribes that do not contain *Alloteropsis*. Using this method, we identified 6657 genes that existed in the common ancestor of *Alloteropsis* and have then been transmitted to some of its descendants. The presence of each of these genes in the five reference genomes was established based on read analyses, as described above for laterally acquired genes. We identified 227 out of these 6657 that were absent from at least one of the five reference genomes, and the presence/absence of each of these 227 genes across the 45 *Alloteropsis* individuals was again established using the read-based analysis.

The rate of losses of these native genes was inferred by mapping their origins on the phylogenetic trees, as described for the

© 2023 The Authors
New Phytologist © 2023 New Phytologist Foundation

14698137, 0, Downloaded from https://nph.onlinelibrary.wiley.com/doi/10.1111/nph.19272 by Test, Wiley Online Library on [09/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

laterally acquired genes. For each gene missing from a reference genome, its loss in the lineage leading to this genome was recorded. As before, if multiple losses were inferred along the branches leading to the reference genome, the gene was not considered when calculating rates. The number of losses across the 100 mapping replicates was then computed per 1 Ma time window. Because losses are not expected to be frequently recovered, the number of observed losses is not expected to decrease with their age, and if the rate of losses is constant, the number of losses per time window should not vary. We consequently computed the rate of losses as the mean number of observed losses across the eight most recent time windows, which represent the time during which *A. angusta* and *A. semialata* evolved separately.

Results

Assembling multiple *Alloteropsis* reference genomes

The size of the *A. semialata* assemblies ranged between 0.62 and 0.86 Gb, which reflects differences in the genome size estimates based on flow cytometry for these diploid accessions (Bianconi *et al.*, 2020), although the mean assembly size was 27% (range 20–32%) smaller than the flow cytometry estimates (Table S3). Accessions RSA5, TAN1 and AUS1 (before super-scaffolding using Dovetail Genomics Chicago and Hi-C data) had similar assembly statistics: mean N50 = 0.18 Mb (SD = 0.01 Mb); mean number of scaffolds = 7059 (SD = 1018); and mean longest scaffold = 1.07 Mb (SD = 0.08 Mb; Table S3). The ZAM1505-10 accession was generally more fragmented than the other three *A. semialata* reference genomes: N50 = 0.07 Mb; number of scaffolds = 19 813; longest scaffold = 0.60 Mb (Table S3). BUSCO analyses indicated that all four *A. semialata* genomes were comparably complete (mean = 89.3%; SD = 2.3%), although the ZAM1505-10 assembly had higher levels of duplication (25.5%) compared with the other three (mean = 7.3%; SD = 1.8%; Table S3). This same pattern was repeated when the BUSCO analysis was performed solely on the genome annotations (Table S3). The increased annotation duplication in ZAM1505-10 is accounted for by our downstream LGT identification pipeline as monophyletic groups of recent duplicates and/or uncollapsed haplotypes are only counted as a single event in the gene tree analyses.

The AANG_UGA4 *A. angusta* assembly was more fragmented and less complete than the four *A. semialata* references, largely owing to the lack of long-read PacBio data for this accession (Table S3). According to the BUSCO analysis, the AANG_UGA4 assembly was only 77.1% complete, and this value decreased to 67.4% when considering the genome annotation. Despite the lower quality of this assembly, it is unlikely to bias our results as it effectively acts as an outgroup to the *A. semialata* accessions.

Widespread lateral gene transfer in *Alloteropsis*

Our phylogenetic pipeline identified an initial 177 laterally acquired genes (both primary (Table S4) and secondary candidates *sensu* Dunning *et al.*, 2019) across the five *Alloteropsis* genomes (Notes S1; Table S5). For 11 of these genes, individual reads could not be reliably assigned to specific gene copies, and these loci were not considered further in this study. For two of the remaining genes, in depth phylogenetic analyses indicated that they had been transferred twice independently to different accessions of *A. semialata* (Figs S2, S3; Notes S1), and the genes resulting from each of these events were considered as independent, leading to a final total of 168 laterally acquired genes (range of 32–100 per genome; Table S5). These genes were acquired from three different subfamilies: 92.2% Panicoideae, 4.8% Chloridoideae and 3.0% Danthonioideae (Fig. 2; Table S6). Within the Panicoideae, the main donors were Cenchrinae ($n = 88$ LGT), Andropogoneae ($n = 54$) and Melinidinae ($n = 7$). For comparison of the laterally acquired genes identified in this and previous studies, see Notes S1 and Table S7.

Genes were acquired as part of fragments of various sizes

Distinct protein-coding laterally acquired genes were assigned to the same genome block if they were adjacent in any of the five reference genomes (Tables S6, S8). In total, the 168 loci could be assigned to 82 different genomic blocks, with the most gene-rich fragment containing 12 distinct genes that spanned over 137 kb in the ZAM1505-10 genome (block 68; Fig. S4; Table S6). A total of 45 laterally acquired genes appeared on their own in the assembled genomes (Table S6). All genes from the same fragments were assigned to the same donor, with one exception (Fig. S5; Table S6). Block 63 contains three genes, and the first two are assigned to Cenchrinae while the last one is assigned to Andropogoneae and is present in more *Alloteropsis* accessions (Figs S5, S6). The third gene is present in three distinct contigs in the assembled genome of TAN1-04B that likely represent post-transfer duplicates, and only one of them is joined with the two other laterally acquired genes (Tables S6, S8). These patterns might result either from a misassembly or from distinct LGT clustering in the genome after independent transfers, as previously suggested in *Hordeum* (Mahelka *et al.*, 2021).

History and geography both explain the distribution of laterally acquired genes

Using a sequencing read analysis, the distribution of each of the 168 laterally acquired genes was established among 40 additional accessions of *Alloteropsis*, including numerous *A. semialata* and *A. angusta* individuals, as well as the outgroup *A. cimicina* (Fig. 1). Together with the five reference genomes, these 45 accessions cover the known geographical and genetic diversity of the genus, with divergence times spanning > 11 Myr (Figs 1, 2). No laterally acquired gene was found in all 45 accessions used in this study. Two were detected in 44 accessions and thus likely acquired before the diversification of *Alloteropsis*, with few other loci shared across species (Fig. 1). A majority of the genes were restricted to phylogenetic subgroups, with five only present in a single individual (ZAM1505-10; Fig. 1; Table S6), suggesting very recent acquisitions. The number shared among pairs

© 2023 The Authors
New Phytologist © 2023 New Phytologist Foundation

14698137, 0 Downloaded from https://nph.onlinelibrary.wiley.com/doi/10.1111/nph.19272 by Test, Wiley Online Library on [09/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

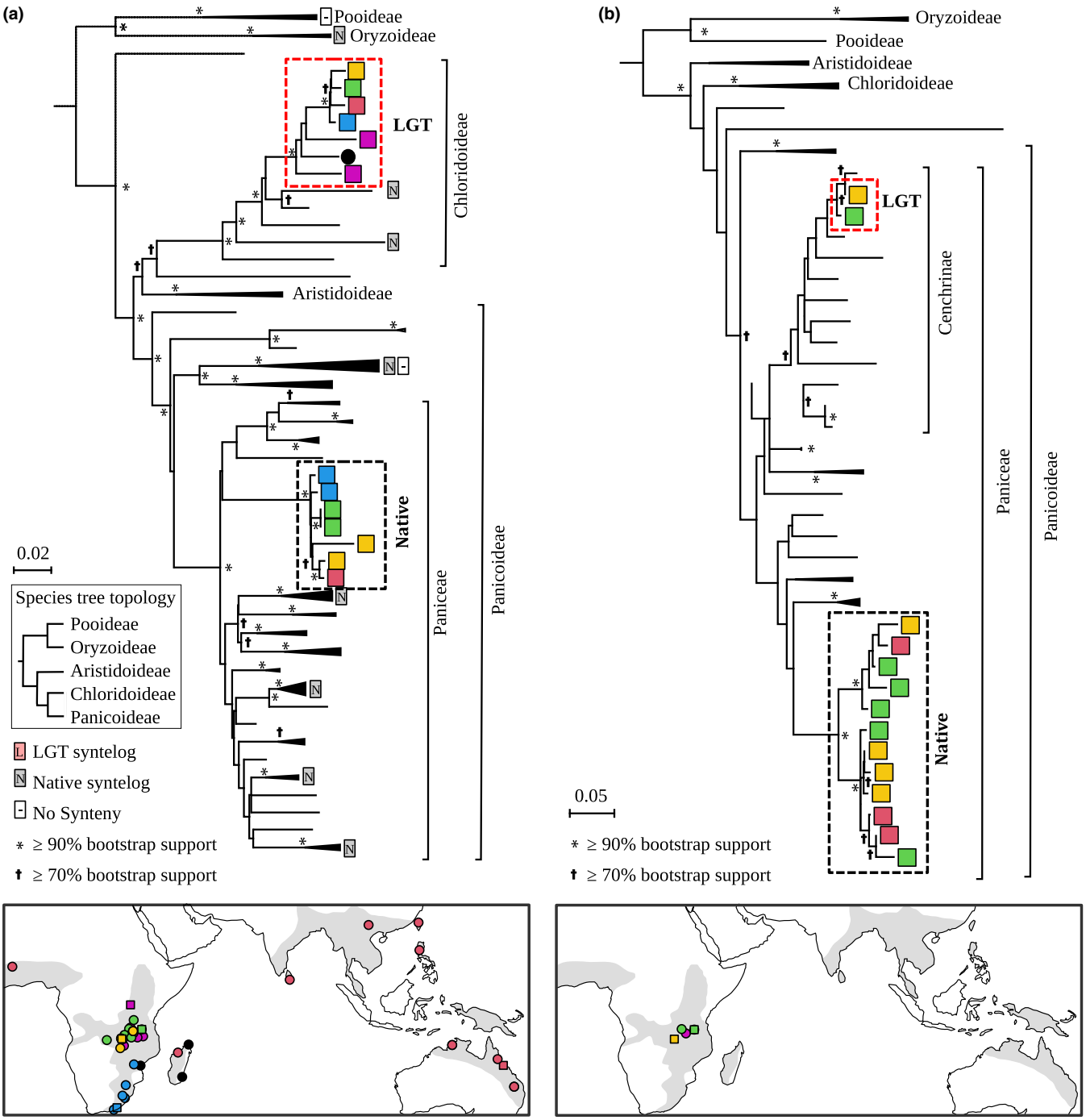

Fig. 2 Phylogenetic evidence and geographical distribution of two exemplar laterally acquired genes. Two examples were selected to represent a laterally acquired gene present in (a) most individuals (LGT-001, reference gene ASEM_ZAM1505-10_31837) and (b) one restricted to a few individuals (LGT-103, reference gene ASEM_TAN1_34267). For each of them, a simplified phylogenetic tree at the top shows the positions of the laterally acquired genes (marked 'LGT') and the native ortholog whose inheritance tracks the species tree. In both groups, the position of the genes extracted from *Alloteropsis cimicina* and the five reference genomes of *Alloteropsis semialata* and *Alloteropsis angusta* is indicated with colours matching (Fig. 1; magenta, AAN-G_UGA4; blue, RSA5-3; green, TAN1-04B; yellow, ZAM1505-10; red, AUS1). The scale represents the number of substitutions per site, and grass subfamilies of interest are shown, and some further divisions within Panicoideae are also shown. For (a) we added synteny information from other model grasses onto the phylogeny that were extracted from a previous analysis of the chromosomal scale AUS1 *A. semialata* genome (Hibdige *et al.*, 2021). Most of the genes in the other model species are syntenic with the native *A. semialata* copy (none are with the laterally acquired gene), confirming that unrecognised parology is not driving the observed phylogenetic patterns. For each laterally acquired gene, the geographical position of accessions bearing them are shown on maps, squares for the reference genomes and circles for resequenced genomes, both coloured by phylogenetic groups as in Fig. 1 (magenta, *A. angusta*; *A. semialata*: blue, Clade I = RSA5-3; green, Clade II = TAN1-04B; yellow, Clade III; red, Clade IV). The grey shading represents the known distribution of *A. semialata*.

© 2023 The Authors
New Phytologist © 2023 New Phytologist Foundation

14698137, 0, Downloaded from https://nph.onlinelibrary.wiley.com/doi/10.1111/nph.19272 by Test, Wiley Online Library on [09/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of individuals decreased with divergence time (Mantel test, $p < 0.0001$, $R^2 = 0.53$; Fig. S7), which supports a gradual accumulation of laterally acquired genes during the diversification of the genus. We conclude that history largely explains the patterns of the distribution of laterally acquired genes within and among *Alloteropsis* species.

Although evolutionary history appears to be the main driver of the present-day distribution of laterally acquired genes, geography also appears to play a role. Some genes present a patchy distribution across the *Alloteropsis* phylogeny, being found in a few accessions belonging to different phylogenetic groups (Fig. 1). These patterns could result from repeated losses following ancient acquisitions, but in some cases likely result from introgression of recently acquired genes. Indeed, some of these laterally acquire genes with a patchy distribution are shared among distantly related accessions, sometimes from different *Alloteropsis* species, which cluster geographically (Fig. 2; Notes S1). Once divergence times were accounted for, the number of laterally acquired loci shared by pairs of accessions decreased with geographical distances (partial Mantel test, $p = 0.0006$, $R^2 = 0.08$; Fig. S7), showing that history and geography both contributed to the present-day distribution of these genes.

The rate of LGT varies among lineages

The transfer and potential loss of each of the 168 laterally acquired genes were individually inferred along one of a hundred time-calibrated phylogenetic trees that best explained the LGT history, using stochastic mapping with 100 pseudoreplicates. The timing of each LGT was then retrieved for the line of ancestors leading to each of the five reference genomes containing it. When more than one transfer was inferred by the stochastic mapping, the gene was discarded (ranging from 25% to 56% of laterally acquired genes; Table S5). The inference of multiple origins might result from repeated losses of older genes or spread of recent ones via introgression, but in both cases, their patchy distribution across the phylogeny means that the LGT event cannot be confidently assigned to a single time point.

As expected, given the random nature of the stochastic mapping approach, the inferred ages of origin varied among the 100 replicates, but there was a consistent increase in observed LGT events towards the present leading to the ZAM1505-10 genome, and to a lesser extent to some of the others (Fig. S8). When considering the average across pseudoreplicates, the number of observed LGTs increased strongly towards the present in RSA5, AUS1 and especially ZAM1505-10, where it fitted closely to an exponential distribution (Fig. 3). Based on fitted exponential decay models, the rates of gains were estimated between 2.66 and 15.74 LGT per Ma for the five genomes. The highest value was observed in ZAM1505-10, with a rate of gains significantly higher than the others, which all had overlapping confidence intervals for this parameter (Fig. 3). If the LGTs for which ages could not be reliably estimated follow the same distribution as the others, the rate of gains would be inflated to between 3.54 and 28.1 LGT per Ma (in ZAM1505-10; Fig. 3).

The rate of losses of LGT per Ma in *A. angusta* was estimated to be 4%, mirroring the weak increase in LGT towards the present (Fig. 3). The modelled relationships indicated that between 11% and 16% of LGTs were lost each Ma in the lineages leading to most *A. semialata*, while 24% were lost in the lineage leading to ZAM1505-10, although the confidence intervals of the latter overlap with those of some others (Fig. 3). The high number of laterally acquired genes present in the genome of ZAM1505-10 therefore results from a higher rate of gains, despite a potentially higher rate of losses (Fig. 3). Overall, these analyses indicate that the genomes of *Alloteropsis* undergo a high turnover of LGT, with repeated gains throughout their history followed by relatively rapid losses. The rates of gains and losses, however, vary among sublineages, creating important variation in laterally acquired gene content, especially within *A. semialata*.

Laterally acquired genes are lost faster than native counterparts

To compare the rate that laterally acquired genes are lost to those that trace the species tree (referred to as native genes), we used a dataset of 6657 relatively conserved genes that were present as a single copy in the common ancestor of *Alloteropsis* and other Paniceae, and that did not have a signal of being acquired through LGT. Using the same sequencing read analyses as for the laterally acquired genes, we estimated the number of native genes lost every Ma, for the lineages leading to each of the five reference genomes. Assuming that the losses whose age could not be estimated are proportionally distributed among the time windows, the fraction of the 6657 genes lost every Ma ranges from 0.25% in *A. angusta* to 0.04–0.06% in the four *A. semialata* individuals (see Notes S1). We conclude that laterally acquired genes are lost up to 500 times faster than native components of the genome of *A. semialata*.

LGT overly contributes to gene presence/absence variation

We evaluated the pangenome variation focusing on the five reference genomes, as genes specific to other accessions cannot be identified based on the data available. Out of 168 laterally acquired genes, only two are present in the five *A. semialata* and *A. angusta* reference genomes and can thus be considered as core genes, with the vast majority (166 genes; 98.8%) appearing as accessory genes. Of the 166 polymorphic laterally acquire genes, 21 are specific to *A. angusta*, while 136 are specific to *A. semialata*, with the other nine shared by some accessions of the two groups (Fig. 4). Of the loci specific to *A. semialata*, 130 vary within *A. semialata*, and most of the pangenome variation was generated by laterally acquired genes restricted to ZAM1505 and to a lesser extent AUS1 (Fig. 4). These results show that recent LGT and subsequent losses create important pangenome variation, within *A. semialata* and among the two sister species. In contrast to genes acquired through LGT, only 3.4% of the 6657 native genes investigated above appear as accessory, and this is mainly due to numerous losses in the *A. angusta* lineages, which contributes in excess to the native pangenome (Fig. 4). This

© 2023 The Authors
New Phytologist © 2023 New Phytologist Foundation

14698137, 0 Downloaded from https://nph.onlinelibrary.wiley.com/doi/10.1111/nph.19272 by Test, Wiley Online Library on [09/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

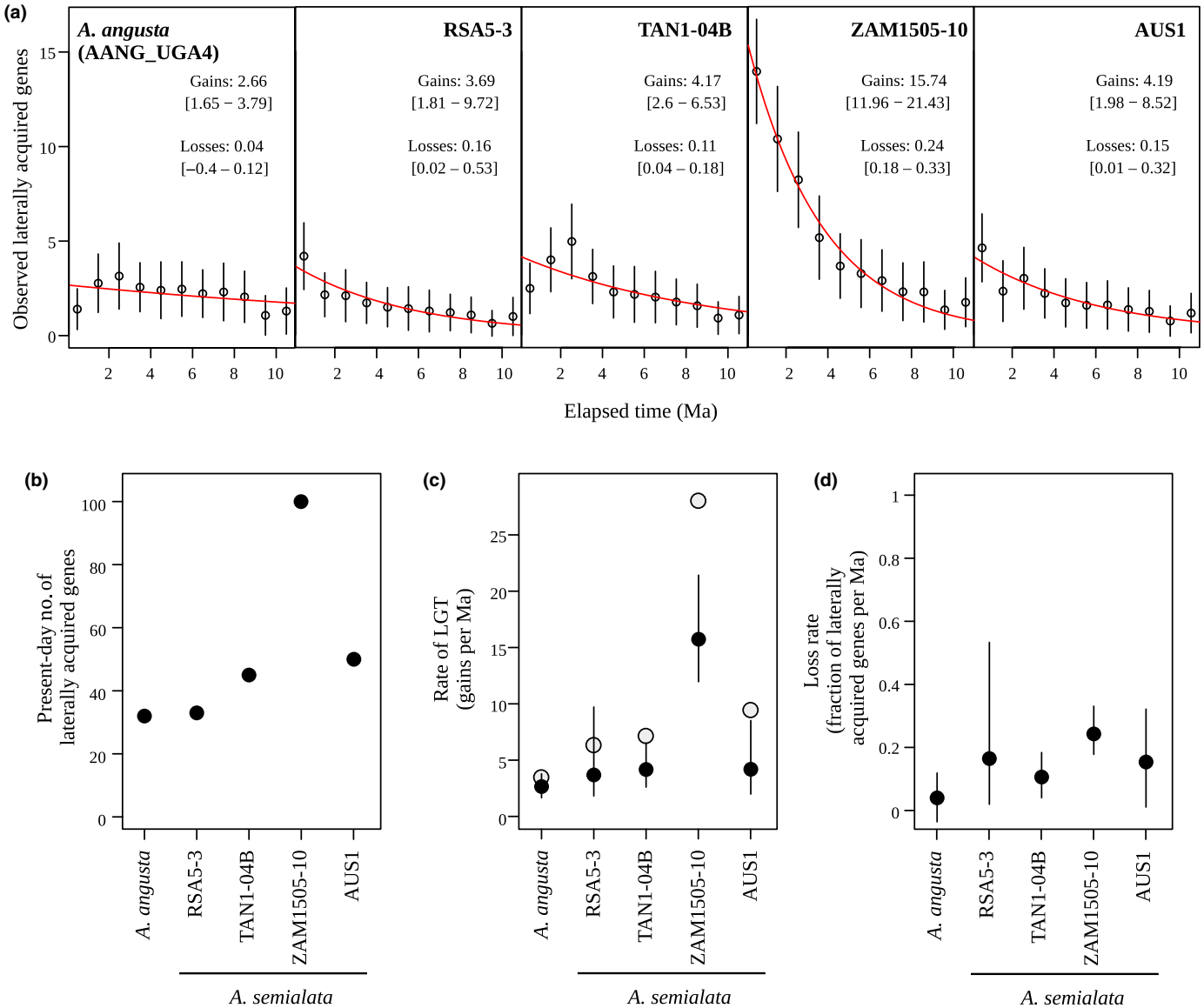

Fig. 3 Rates of lateral gene transfer and subsequent loss of laterally acquired genes. (a) Numbers of lateral gene transfers (LGT) assigned to different 1 Ma time slices. For each of the five *Alloteropsis* reference genomes, named at the top, the points indicate the mean and the bars the 95% intervals across 100 pseudoreplicates. The red curves show the fitted exponential decay model. The inferred rates of LGT and the subsequent loss of the laterally acquire genes per Ma are indicated, with 95% confidence intervals in square brackets. (b) The number of LGT observed in the genome of each of the five reference genomes is indicated. (c) The inferred rate of LGT gains per Ma is shown with closed circles for the five references genomes, with bars showing the 95% intervals across replicates. The open circles show the estimates obtained assuming that laterally acquire genes for which a single time of origin could not be estimated are distributed through time proportionally to those with estimated ages. (d) The inferred rate of laterally acquired gene loss per Ma is indicated for the five reference genomes. The points show the mean and the bars the 95% intervals across 100 pseudoreplicates.

result is not due to the more fragmented and incomplete nature of the *A. angusta* reference genome as this analysis was performed on the unassembled reads. Within *A. semialata*, the four accessions contribute in similar proportions to the native pangenome variation, which stands in stark contrast to the laterally acquired gene pangenome contributions (Fig. 4).

Our analysis of native genes was conservative, and we consequently captured as few as one-tenth of all annotated protein-coding genes (Table S1). If the proportion of variable genes was maintained, we would thus expect up to 2270 accessory native genes. This is a lot lower than the number of accessory genes identified in other grass pangenomes, such as maize where 69% are variable (Hufford *et al.*, 2021). However, a majority of these variable genes are only found in maize (Hufford *et al.*, 2021), meaning they would not be included in our conservative phylogenetic approach. It is also worth noting that our analyses are based on read presence/absence and not gene annotation presence/absence. This means that our approach is again conservative as it is not affected by genome assembly and annotation quality, and it also counts pseudogenes still present in the genome. While these methodological and data volume differences (> 5× fewer *Alloteropsis* than maize assemblies) result in a

© 2023 The Authors
New Phytologist © 2023 New Phytologist Foundation

14698137, 0, Downloaded from https://nph.onlinelibrary.wiley.com/doi/10.1111/nph.19272 by Test, Wiley Online Library on [09/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

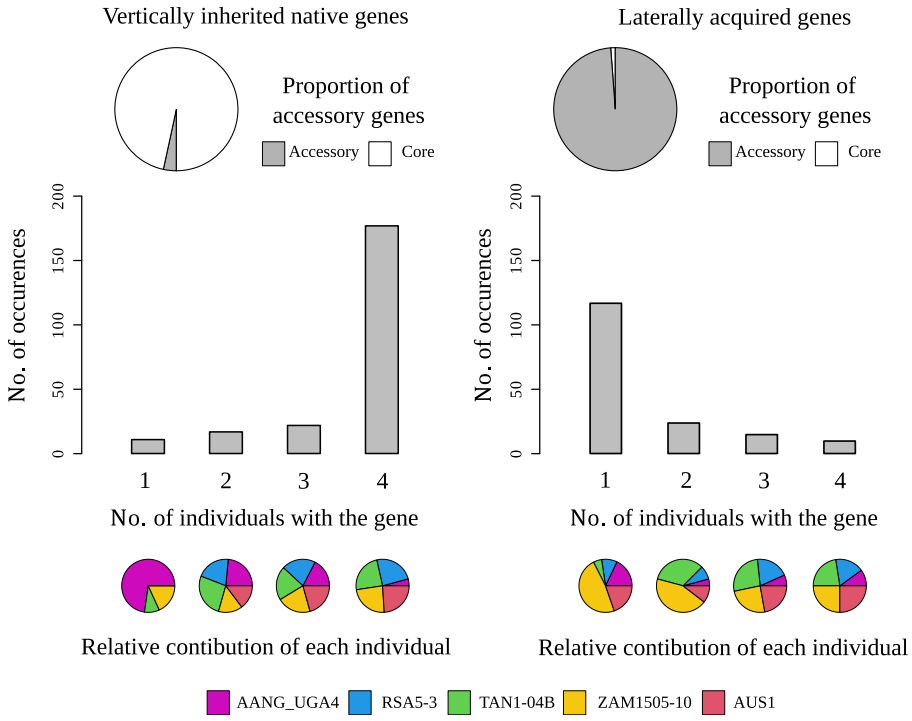

Fig. 4 Contribution to pangenome variation in *Alloteropsis*. Pie charts at the top show the proportion of accessory (presence variable) and core (presence fixed) genes for native loci inherited vertically from the ancestors of *Alloteropsis*, and those resulting from lateral gene transfer. The bar charts below indicate the number of reference genomes each accessory gene is found in, and the pie charts below that show the proportions found in each reference genome.

fewer number of variable genes in *Alloteropsis* compared with other species, our pipelines to trace the evolutionary history of laterally acquired and native genes is equally conservative and is therefore comparable.

In total, 99% of detected LGT in *Alloteropsis* are variable, which equates to 166 genes in total. The number of laterally acquired genes identified is also likely to be underestimated, but they still represent at least 7% of the variable genes, despite accounting for < 0.5% of all protein-coding genes. Overall, these patterns indicate that, despite their low relative numbers, the laterally acquired genes disproportionately contribute to the variable portion of the *Alloteropsis* pangenome, and that this excess contribution stems from both a high rate of LGT and secondary losses.

Functions of the laterally acquire genes

The final list of 168 laterally acquired genes was annotated against the SwissProt database, revealing a diverse set of functions, including genes associated with C_4 photosynthesis, disease resistance and abiotic stress tolerance (Table S9). A subsequent GO enrichment analysis identified four significantly (adjusted $P < 0.05$) overrepresented categories: cellulose biosynthetic process, cellulose synthase (UDP-forming) activity, pre-mRNA 3′-splice site binding and ribonuclease P activity (Table S10). It is unsurprising that there are relatively few enriched categories when all are considered at once, as they include both the genes that were potentially selected for in addition to adjacent loci that hitchhiked along as part of the same fragment of DNA (Olofsson *et al.*, 2019). Further work to disentangle the two may provide clarity as to whether genes with certain functions are preferentially retained after the initial transfer.

Discussion

Continual birth and death of laterally acquired genes

We can only detect laterally acquired genes that are retained until the present. Those detected therefore likely represent the tip of the iceberg of a much higher background rate of transfer and subsequent loss through drift or negative selection. Comparison of the laterally acquired gene content among multiple accessions of the same group can help infer the dynamics of gene transfers through time. For instance, those restricted to accessions of *A. semialata* from Australia presenting a high similarity with populations of the donor species from the same region must have been acquired after the colonization of Australia, inferred in the last 0.5 Ma (Olofsson *et al.*, 2019). Conversely, genes present in all *Alloteropsis* species which accumulated substitutions after the transfer (Fig. 2) were likely acquired before the diversification of the genus some 11 Ma. The exact timing of a majority of LGTs is difficult to precisely date, but their distribution among accessions can be used to model the temporal spread of acquisitions, but also of subsequent losses.

The laterally acquired genes present in a given genome represent those that were gained and persisted minus the proportion that have been lost. Assuming that the rate of gain, corresponding to the rate of integration of foreign genes in the genome of *A. semialata*, and the rate of subsequent losses are constant through time, the number of genes persisting at present should decrease with their age, following an exponential decay model. Such patterns were indeed observed for those found in the genomes of the *A. semialata* accessions (Fig. 3). These analyses bring direct support for the accumulation of laterally acquired genes in

© 2023 The Authors
New Phytologist © 2023 New Phytologist Foundation

14698137, 0 Downloaded from https://nph.onlinelibrary.wiley.com/doi/10.1111/nph.19272 by Test, Wiley Online Library on [09/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the genomes of plants and also indicate that between 6 and 28 loci entered the genomic lineages of the four *A. semialata* analysed here every million years (Fig. 3). However, between 11% and 24% of those that were originally integrated are lost every million years, a proportion that vastly exceeds the fraction of native genes lost every million years (0.04–0.06% for *A. semialata*; Fig. S9). These numbers indicate a high turnover rate of laterally acquired genes, with potentially > 20 foreign genes entering the recipient gene pool every million years, but half of those being lost after 3–6 Ma. Why some genes are retained for long periods and others are rapidly lost is presently unknown, although it is likely to reflect whether the selection pressure to retain these loci is temporal or sustained. For example, laterally acquired genes associated with disease resistance could rapidly spread through a population during an epidemic, but then be lost through drift as chances of infection subside. Conversely, genes incorporated into key pathways may be retained over the long term, such as copies of core C_4 photosynthetic genes acquired by *A. semialata* that have functionally replaced the native versions that have become pseudogenes (Phansopa *et al.*, 2020).

Lineage-specific dynamics

While our models assume constant rates of LGT gains and subsequent losses through time, this assumption is likely violated. First, the raw number of DNA transfers is likely to vary as transfer opportunities fluctuate. Such variation might represent changes in the phenotype of the recipient. For instance, high levels of selfing would be expected to reduce LGT via illegitimate pollination, whereas increased vegetative propagation would provide more opportunities for LGT through root inosculation. In addition, the probability of DNA transfers likely depends on the presence of potential donors in close contact, which can vary as species migrate. Second, the probability that transferred DNA persists over generations depends on the demography of the recipient species. Large effective population sizes would increase the chance of advantageous laterally acquired genes spreading under selection, while increased drift in small populations would increase the retention of neutral loci. The causal factors are difficult to identify, but such variation is likely responsible for both deviations from the models within each lineage and differences among them. It is also worth noting that the rates we present are almost certainly underestimated due to technical limitations. This includes missing genes that were not assembled/annotated in the reference genomes (*c.* 10% of loci in each *A. semialata* reference genome based on the BUSCO analysis), in addition to our conservative detection pipeline which excludes genes which are not present in enough species to infer robust phylogenetic trees (Hibdige *et al.*, 2021). However, what is important is that these issues are not restricted to any particular *A. semilata* lineage meaning the inferred rates can be compared.

The most striking difference concerns the lineage of ZAM1505, which accumulated significantly more laterally acquired genes than the other lineages (Fig. 3). This accession comes from a region of Zambia where *A. semialata* was often found forming multispecies clumps with some known donors (Fig. S10), potentially providing increased opportunities for LGT. In addition, this accession is located near the inferred centre of origin of the species (Bianconi *et al.*, 2020), and constant and large population sizes might have favoured the integration of beneficial genes, as well as physically linked neutral loci as reported in Australia (Olofsson *et al.*, 2019). These scenarios remain speculative, but the patterns reported here show that some lineages overly contribute to the laterally acquired gene content of a species.

The precise mechanism behind the transfers is currently unknown, although many have been proposed (Dunning *et al.*, 2019; Christin *et al.*, 2012; Hibdige *et al.*, 2021). Based on current evidence, the most mechanisms likely involve reproductive contamination through illegitimate pollination (Christin *et al.*, 2012; Pereira *et al.*, 2022). This would potentially mirror plant transformation techniques such as repeated pollination (Shan *et al.*, 2005) or pollen tube pathway-mediated transformation (Ali *et al.*, 2015) where the reproductive process is effectively contaminated with DNA from a third individual. These transformation methods require minimal human intervention and could therefore occur naturally in wind-pollinated species, driving the observed grass-to-grass LGT through reproductive contamination (Pereira *et al.*, 2022).

LGT excessively contributes to pangenome variation

We show here that LGT, which is responsible for the acquisition of < 1% of all genes present in a given accession of *Alloteropsis*, excessively contribute to both the pangenome of *A. semialata* and the joint pangenome of *A. semialata* and *A. angusta* (Fig. 4). Because they are continuously acquired and then lost more rapidly than native genes (Figs 3, S8), most laterally acquired genes are indeed variable, within the species and even within some populations (Fig. S6; Olofsson *et al.*, 2016). Our results moreover show that even if genes are preferentially acquired by a given sublineage, such as ZAM1505, they can greatly contribute to the species-level pangenome (Fig. 4) and might later be introgressed to other populations (Fig. 1; Olofsson *et al.*, 2016). Indeed, the different sublineages of *A. semialata* and even the two sister species *A. angusta* and *A. cimicina* occasionally undergo gene flow (Olofsson *et al.*, 2016; Curran *et al.*, 2022) so that the standing variation created by LGT has the potential to fuel adaptation throughout the group.

Unlike duplicates of native genes, genes acquired through LGT have diverged from the other genes in the genome for at least as long as the divergence time between the donor and the recipient, which in the case of examples reported here extends to > 40 Ma (Christin *et al.*, 2008). These laterally acquired genes consequently add diversity to the recipient genomes, both in terms of expression patterns (Dunning *et al.*, 2019) and coding sequences affecting the catalytic properties of the encoded enzymes (Phansopa *et al.*, 2020). We therefore conclude that the high turnover of laterally acquired genes revealed here creates important pangenome variation and therefore impacts the evolutionary potential of species undergoing such DNA exchanges.

© 2023 The Authors
New Phytologist © 2023 New Phytologist Foundation

14698137, 0, Downloaded from https://nph.onlinelibrary.wiley.com/doi/10.1111/nph.19272 by Test, Wiley Online Library on [09/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

14698137, 0, Downloaded from https://nph.onlinelibrary.wiley.com/doi/10.1111/nph.19272 by Test, Wiley Online Library on [09/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Conclusions

Grasses appear to frequently undergo lateral gene transfer. Here, we detect laterally acquired genes in five *de novo* reference genomes belonging to genetically divergent sublineages within the grass *A. semialata* and its sister species *A. angusta*. We identify a total of 168 laterally acquired genes, but only two are shared by all five genomes, and the distribution of these loci among 45 *Alloteropsis* individuals suggests a few old acquisitions and many recent ones that are restricted to sublineages. Analyses of their distribution among individuals in a phylogenetic context allowed estimates of the rates of gains and subsequent losses, using an exponential decay model. We estimated that up to 28 LGT per Ma were accumulated by one lineage of *A. semialata*, with up to one-quarter of the acquired genes subsequently lost every million years. The rate of LGT varied drastically among the five accessions, potentially reflecting differences in opportunities for transfers to occur. This high turnover created important inter- and intraspecific variation in laterally acquired gene content, with almost all of them being polymorphic, compared with only 3.4% of native genes. LGT therefore excessively contributes to the pangenome variation in this group. Because the laterally acquired genes provide novelty to the recipient genomes and can be subsequently introgressed among related species, the standing variation revealed here has the potential to fuel rapid adaptation in these grasses.

Acknowledgements

This work was funded by the Natural Environment Research Council grant NE/V000012/1, PAC is funded by a Royal Society University Research Fellowship (grant URF/R/180022), and LTD is funded by a NERC fellowship (grant NE/T011025/1).

Competing interests

None declared.

Author contributions

PR, CP, P-AC and LTD designed the study. PR, MEB and LTD generated the genome data. MEB and LTD assembled and annotated the genomes. PR, LP, P-AC and LTD analysed the data. All authors interpreted the results and helped write the manuscript.

ORCID

Matheus E. Bianconi https://orcid.org/0000-0002-1585-5947
Pascal-Antoine Christin https://orcid.org/0000-0001-6292-8734
Luke T. Dunning https://orcid.org/0000-0002-4776-9568
Christian Parisod https://orcid.org/0000-0001-8798-0897
Lara Pereira https://orcid.org/0000-0001-5184-8587
Pauline Raimondeau https://orcid.org/0000-0002-1005-2926

Data availability

Raw sequence data, genome assemblies and annotations generated as part of this study are available from NCBI GenBank under BioProject PRJNA824797. All phylogenies and alignments have been made available on dryad: https://doi.org/10.5061/dryad.sbcc2fr9s, where assemblies and annotations have also been made available. The scripts used in this study are available from GitHub: https://github.com/Sheffield-Plant-Evolutionary-Genomics/panLGT-Alloteropsis-2022.

References

Ali A, Bang SW, Chung SM, Staub JE. 2015. Plant transformation via pollen tube-mediated gene transfer. *Plant Molecular Biology Reporter* **33**: 742–747.

Bianconi ME, Dunning LT, Curran EV, Hidalgo O, Powell RF, Mian S, Leitch IJ, Lundgren MR, Manzi S, Vorontsova MS *et al.* 2020. Contrasted histories of organelle and nuclear genomes underlying physiological diversification in a grass species. *Proceedings of the Royal Society B* **287**: 20201960.

Bouckaert R, Vaughan TG, Barido-Sottani J, Duchêne S, Fourment M, Gavryushkina A, Heled J, Jones G, Kühnert D, De Maio N *et al.* 2019. BEAST 2.5: an advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology* **15**: e1006650.

Brockhurst MA, Harrison E, Hall JP, Richards T, McNally A, MacLean C. 2019. The ecology and evolution of pangenomes. *Current Biology* **29**: R1094–R1103.

Cai L, Arnold BJ, Xi Z, Khost DE, Patel N, Hartmann CB, Manickam S, Sasirat S, Nikolov LA, Mathews S *et al.* 2021. Deeply altered genome architecture in the endoparasitic flowering plant *Sapria himalayana* Griff. (Rafflesiaceae). *Current Biology* **31**: 1002–1011.

Christin PA, Besnard G, Samaritani E, Duvall MR, Hodkinson TR, Savolainen V, Salamin N. 2008. Oligocene CO_2 decline promoted C_4 photosynthesis in grasses. *Current Biology* **18**: 37–43.

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

Curran EV, Scott MS, Olofsson JK, Nyirenda F, Sotelo G, Bianconi ME, Manzi S, Besnard G, Pereira L, Christin PA. 2022. Hybridization boosts dispersal of two contrasted ecotypes in a grass species. *Proceedings of the Royal Society B* **289**: 20212491.

Dorrell RG, Villain A, Perez-Lamarque B, Audren de Kerdrel G, McCallum G, Watson AK, Ait-Mohamed O, Alberti A, Corre E, Frischkorn KR *et al.* 2021. Phylogenomic fingerprinting of tempo and functions of horizontal gene transfer within ochrophytes. *Proceedings of the National Academy of Sciences, USA* **118**: e2009974118.

Dunning LT, Olofsson JK, Parisod C, Choudhury RR, Moreno-Villena JJ, Yang Y, Dionora J, Quick WP, Park M, Bennetzen JL *et al.* 2019. Lateral transfers of large DNA fragments spread functional genes among grasses. *Proceedings of the National Academy of Sciences, USA* **116**: 4416–4425.

El Baidouri M, Carpentier MC, Cooke R, Gao D, Lasserre E, Llauro C, Mirouze M, Picault N, Jackson SA, Panaud O. 2014. Widespread and frequent horizontal transfers of transposable elements in plants. *Genome Research* **24**: 831–838.

Fan X, Qiu H, Han W, Wang Y, Xu D, Zhang X, Bhattacharya D, Ye N. 2020. Phytoplankton pangenome reveals extensive prokaryotic horizontal gene transfer of diverse functions. *Science Advances* **6**: eaba0111.

Fernández R, Gabaldón T. 2020. Gene gain and loss across the metazoan tree of life. *Nature Ecology & Evolution* **4**: 524–533.

Golicz AA, Bayer PE, Bhalla PL, Batley J, Edwards D. 2020. Pangenomics comes of age: from bacteria to plant and animal applications. *Trends in Genetics* **36**: 132–145.

Gordon SP, Contreras-Moreira B, Woods DP, Des Marais DL, Burgess D, Shu S, Stritt C, Roulin AC, Schackwitz W, Tyler L *et al.* 2017. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. *Nature Communications* **8**: 1–13.

© 2023 The Authors
New Phytologist © 2023 New Phytologist Foundation

Han DY, Han PJ, Rumbold K, Koricha AD, Duan SF, Song L, Shi JY, Li K, Wang QM, Bai FY. 2021. Adaptive gene content and allele distribution variations in the wild and domesticated populations of *Saccharomyces cerevisiae*. *Frontiers in Microbiology* **12**: 247.

Hibdige SG, Raimondeau P, Christin PA, Dunning LT. 2021. Widespread lateral gene transfer among grasses. *New Phytologist* **230**: 2474–2486.

Hufford MB, Seetharam AS, Woodhouse MR, Chougule KM, Ou S, Liu J, Ricci WA, Guo T, Olson A, Qiu Y *et al.* 2021. *De novo* assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science* **373**: 655–662.

Kado T, Innan H. 2018. Horizontal gene transfer in five parasite plant species in Orobanchaceae. *Genome Biology and Evolution* **10**: 3196–3210.

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**: 605–618.

Lefort V, Longueville JE, Gascuel O. 2017. SMS: smart model selection in PHYML. *Molecular Biology and Evolution* **34**: 2422–2424.

Li FW, Brouwer P, Carretero-Paulet L, Cheng S, De Vries J, Delaux PM, Eily A, Koppers N, Kuo LY, Li Z *et al.* 2018. Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nature Plants* **4**: 460–472.

Li FW, Villarreal JC, Kelly S, Rothfels CJ, Melkonian M, Frangedakis E, Ruhsam M, Sigel EM, Der JP, Pittermann J *et al.* 2014. Horizontal transfer of an adaptive chimeric photoreceptor from bryophytes to ferns. *Proceedings of the National Academy of Sciences, USA* **111**: 6672–6677.

Ma J, Wang S, Zhu X, Sun G, Chang G, Li L, Hu X, Zhang S, Zhou Y, Song CP *et al.* 2022. Major episodes of horizontal gene transfer drove the evolution of land plants. *Molecular Plant* **15**: 857–871.

Mahelka V, Krak K, Fehrer J, Caklová P, Nagy Nejedla M, Čegan R, Kopecký D, 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**: 1141–1164.

Mahelka V, Krak K, Kopecký D, Fehrer J, Šafář J, Bartoš J, Hobza R, Blavet N, Blattner FR. 2017. Multiple horizontal transfers of nuclear ribosomal genes between phylogenetically distinct grass lineages. *Proceedings of the National Academy of Sciences, USA* **114**: 1726–1731.

Murat F, Armero A, Pont C, Klopp C, Salse J. 2017. Reconstructing the genome of the most recent common ancestor of flowering plants. *Nature Genetics* **49**: 490–496.

Olofsson JK, Bianconi M, Besnard G, Dunning LT, Lundgren MR, Holota H, Vorontsova MS, Hidalgo O, Leitch IJ, Nosil P *et al.* 2016. Genome biogeography reveals the intraspecific spread of adaptive mutations for a complex trait. *Molecular Ecology* **25**: 6107–6123.

Olofsson JK, Dunning LT, Lundgren MR, Barton HJ, Thompson J, Cuff N, Ariyarathne M, Yakandawala D, Sotelo G, Zeng K *et al.* 2019. Population-specific selection on standing variation generated by lateral gene transfers in a grass. *Current Biology* **29**: 3921–3927.

Paradis E, Schliep K. 2019. APE 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**: 526–528.

Pereira L, Christin PA, Dunning LT. 2022. The mechanisms underpinning lateral gene transfer between grasses. *Plants, People, Planet* **5**: 1–11.

Phansopa C, Dunning LT, Reid JD, Christin PA. 2020. Lateral gene transfer acts as an evolutionary shortcut to efficient C_4 biochemistry. *Molecular Biology and Evolution* **37**: 3094–3104.

Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin EV. 2014. Genomes in turmoil: quantification of genome dynamics in prokaryote supergenomes. *BMC Biology* **12**: 1–19.

Revell LJ. 2012. PHYTOOLS: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution* **3**: 217–223.

Reynolds HT, Vijayakumar V, Gluck-Thaler E, Korotkin HB, Matheny PB, Slot JC. 2018. Horizontal gene cluster transfer increased hallucinogenic mushroom diversity. *Evolution Letters* **2**: 88–101.

Schlötterer C. 2015. Genes from scratch–the evolutionary fate of *de novo* genes. *Trends in Genetics* **31**: 215–219.

Shan X, Liu Z, Dong Z, Wang Y, Chen Y, Lin X, Long L, Han F, Dong Y, Liu B. 2005. Mobilization of the active MITE transposons *mPing* and *Pong* in rice by introgression from wild rice (*Zizania latifolia* Griseb.). *Molecular Biology and Evolution* **22**: 976–990.

Soanes D, Richards TA. 2014. Horizontal gene transfer in eukaryotic plant pathogens. *Annual Review of Phytopathology* **52**: 583–614.

Tao Y, Luo H, Xu J, Cruickshank A, Zhao X, Teng F, Hathorn A, Wu X, Liu Y, Shatte T *et al.* 2021. Extensive variation within the pan-genome of cultivated and wild sorghum. *Nature Plants* **7**: 766–773.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS *et al.* 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences, USA* **102**: 13950–13955.

Van Etten J, Bhattacharya D. 2020. Horizontal gene transfer in eukaryotes: not if, but how much? *Trends in Genetics* **36**: 915–925.

Vavrek MJ. 2011. FOSSIL: palaeoecological and palaeogeographical analysis tools. *Palaeontologia Electronica* **14**: 16.

Wang H, Sun S, Ge W, Zhao L, Hou B, Wang K, Lyu Z, Chen L, Xu S, Guo J *et al.* 2020. Horizontal gene transfer of *Fhb7* from fungus underlies *Fusarium* head blight resistance in wheat. *Science* **368**: eaba5435.

Wickell DA, Li FW. 2020. On the evolutionary significance of horizontal gene transfers in plants. *New Phytologist* **225**: 113–117.

Wu D, Jiang B, Ye CY, Timko MP, Fan L. 2022. Horizontal transfer and evolution of the biosynthetic gene cluster for benzoxazinoid in plants. *Plant Communications* **3**: 100320.

Xia J, Guo Z, Yang Z, Han H, Wang S, Xu H, Yang X, Yang F, Wu Q, Xie W *et al.* 2021. Whitefly hijacks a plant detoxification gene that neutralizes plant toxins. *Cell* **184**: 1693–1705.

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

Supporting Information

Additional Supporting Information may be found online in the Supporting Information section at the end of the article.

Fig. S1 Schematic of the primary laterally acquired gene identification pipeline.

Fig. S2 Phylogenetic tree of gene ZAM1505-10-04117 and homologues.

Fig. S3 Phylogenetic tree of gene ZAM1505-10-42 046 and homologues.

Fig. S4 Most gene-rich laterally acquired fragment.

Fig. S5 Laterally acquired fragment containing genes from multiple donors.

Fig. S6 Distribution of laterally acquired genomic blocks.

Fig. S7 Effects of history and geography on the distribution of laterally acquired genes among accessions.

Fig. S8 Patterns of gains of laterally acquired genes through time for pseudoreplicates.

Fig. S9 Losses of native genes through time.

Fig. S10 Examples of multispecies clumps in Zambia.

© 2023 The Authors
New Phytologist © 2023 New Phytologist Foundation

14698137, 0, Downloaded from https://nph.onlinelibrary.wiley.com/doi/10.1111/nph.19272 by Test, Wiley Online Library on [09/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Methods S1 Additional methods including genome sequencing (assembly and annotation), identification of laterally acquired genes, gene ontology enrichment and molecular dating.

Notes S1 Additional results including detection of laterally acquired genes, donor identification, phylogenetic distribution, detecting multiple transfers, the role of introgression and comparing the rates of gene loss.

Table S1 List of cleaned sequence datasets used for genome analyses.

Table S2 Datasets used in tree construction.

Table S3 Genome assembly statistics for the accessions of *Alloteropsis* used in this study.

Table S4 Detailed information on primary lateral gene transfer candidates.

Table S5 Summary of lateral gene transfer numbers per genome.

Table S6 Properties and distribution of detected lateral gene transfer.

Table S7 Comparison of lateral gene transfer detected from the AUS1 genome assembly in different studies.

Table S8 Detected lateral gene transfer and number of reads assigned to each of them.

Table S9 SwissProt lateral gene transfer annotations.

Table S10 Gene ontology enrichment analysis results.

Please note: Wiley is not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

© 2023 The Authors
New Phytologist © 2023 New Phytologist Foundation

14698137, 0 Downloaded from https://nph.onlinelibrary.wiley.com/doi/10.1111/nph.19272 by Test, Wiley Online Library on [09/10/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License