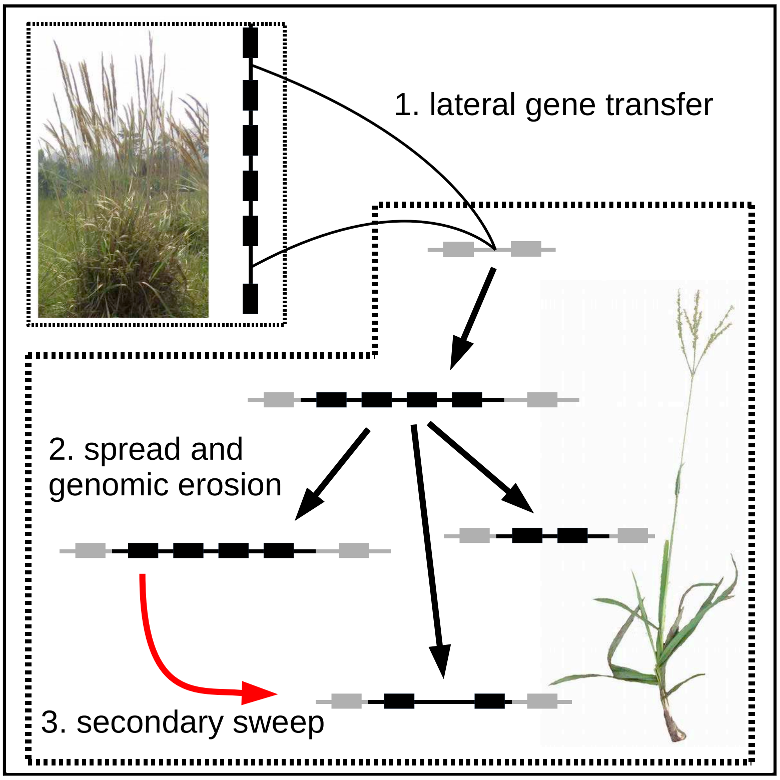
Current Biology

Report

# Population-Specific Selection on Standing Variation Generated by Lateral Gene Transfers in a Grass

## Graphical Abstract



## Authors

Jill K. Olofsson, Luke T. Dunning,
Marjorie R. Lundgren, ...,
Colin P. Osborne, Patrik Nosil,
Pascal-Antoine Christin

## Correspondence

p.christin@sheffield.ac.uk

## In Brief

Olofsson et al. demonstrate that laterally acquired genomic fragments rapidly spread among established populations of a grass, but subsequent genomic erosion creates polymorphisms for some neutral hitchhikers. These hitchhikers can be involved in secondary sweeps, showing that lateral gene transfers have delayed adaptive impacts.

## Highlights

- Laterally acquired genes rapidly spread among established populations of a grass
- Subsequent genomic erosion created neutral gene presence-absence polymorphisms
- One of these neutral genes was secondarily swept into a population
- Lateral gene transfers have both direct and delayed adaptive impacts

Olofsson et al., 2019, Current Biology *29*, 1–7
November 18, 2019 © 2019 The Author(s). Published by Elsevier Ltd.
https://doi.org/10.1016/j.cub.2019.09.023

Current Biology
Report

CellPress

# Population-Specific Selection on Standing Variation Generated by Lateral Gene Transfers in a Grass

**Jill K. Olofsson,[1] Luke T. Dunning,[1] Marjorie R. Lundgren,[1,5] Henry J. Barton,[1,6] John Thompson,[2] Nicholas Cuff,[3] Menaka Ariyarathne,[4] Deepthi Yakandawala,[4] Graciela Sotelo,[1] Kai Zeng,[1] Colin P. Osborne,[1] Patrik Nosil,[1,7] and Pascal-Antoine Christin[1,8,*]**

[1]Department of Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, UK
[2]Queensland Herbarium, Department of Science, Information Technology and Innovation (DSITI), Mt Cooth-tha Botanic Gardens, Toowong, QLD 4066, Australia
[3]Northern Territory Herbarium, Department of Environment and Natural Resources, PO Box 496, Palmerston, NT 0831, Australia
[4]Department of Botany, Faculty of Science, University of Peradeniya, Galaha Road, Peradeniya 20400, Sri Lanka
[5]Present address: Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YW, UK
[6]Present address: Organismal and Evolutionary Biology Research Programme, University of Helsinki, Viikinkaari 9 (PL 56), Helsinki FI-00014, Finland
[7]Present address: Centre for Functional Ecology and Evolution, National Centre for Scientific Research, 1919 route de Mende, Montpellier 34000, France
[8]Lead Contact
*Correspondence: p.christin@sheffield.ac.uk
https://doi.org/10.1016/j.cub.2019.09.023

## SUMMARY

**Evidence of eukaryote-to-eukaryote lateral gene transfer (LGT) has accumulated in recent years [1–14], but the selective pressures governing the evolutionary fate of these genes within recipient species remain largely unexplored [15, 16]. Among nonparasitic plants, successful LGT has been reported between different grass species [5, 8, 11, 16–19]. Here, we use the grass *Alloteropsis semialata*, a species that possesses multigene LGT fragments that were acquired recently from distantly related grass species [5, 11, 16], to test the hypothesis that the successful LGT conferred an advantage and were thus rapidly swept into the recipient species. Combining whole-genome and population-level RAD sequencing, we show that the multigene LGT fragments were rapidly integrated in the recipient genome, likely due to positive selection for genes encoding proteins that added novel functions. These fragments also contained physically linked hitchhiking protein-coding genes, and subsequent genomic erosion has generated gene presence-absence polymorphisms that persist in multiple geographic locations, becoming part of the standing genetic variation. Importantly, one of the hitchhiking genes underwent a secondary rapid spread in some populations. This shows that eukaryotic LGT can have a delayed impact, contributing to local adaptation and intraspecific ecological diversification. Therefore, while short-term LGT integration is mediated by positive selection on some of the transferred genes, physically linked hitchhikers can remain functional and augment the standing genetic variation with delayed adaptive consequences.**

## RESULTS AND DISCUSSION

### Strong Population Structure following a Single Colonization of Oceania

A previous in-depth analysis of the genome of an Australian accession of the paleotropical grass *Alloteropsis semialata* revealed that its genome is scattered with at least 23 fragments of DNA laterally acquired from other grasses [11]. These fragments encompass a total of 59 protein-coding genes, some of which are expressed, while others are pseudogenized [11]. Some of these LGT fragments are geographically restricted, with three fragments containing multiple genes detected only in Australia and the Philippines (fragments LGT_A, LGT_B, and LGT_C), the latter resulting from a putative admixture event from Australia [11]. These fragments were likely acquired by lineages that spread to Oceania from Southeast Asia within the last million years [20], providing tractable systems to test the hypothesis that positive selection led to the rapid spread of LGT among members of the recipient species.

The evolutionary history and population structure of *A. semialata* within Southeast Asia and Oceania was first established to determine the conditions in which the LGT occurred (Figure 1). Whole-genome data for 11 individuals of *A. semialata* distributed across Asia and Oceania were generated to supplement existing data for this species [11]. Phylogenetic relationships inferred from plastid and mitochondrial genomes support a single Asia-Oceania clade, in which a nested monophyletic Oceanian clade is sister to accessions from Sri Lanka and Thailand (Figure S1). The seed-transported organelles therefore indicate a single colonization of Asia-Oceania,

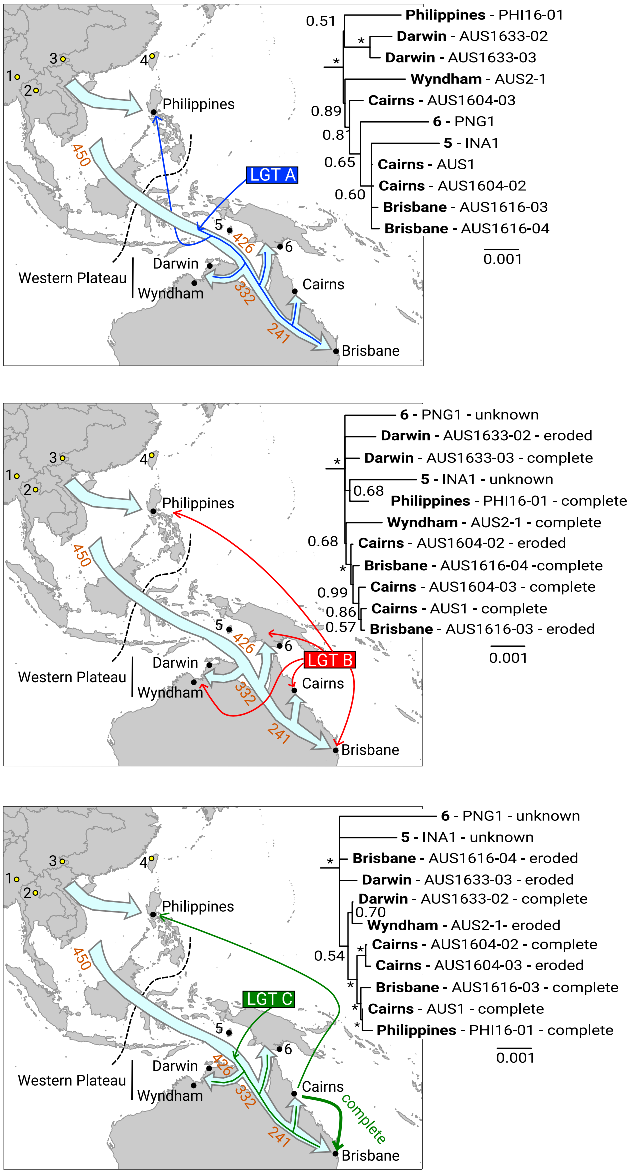


**Figure 1. Evolutionary History of Three Laterally Acquired Fragments**

For each of the three fragments containing multiple genes, its acquisition and spread are shown with dark colored lines on top of the colonization history (light blue wide arrows, inferred from organelle phylogenies; Figure S1). Arrows independent of the colonization history represent putative introgression events. The locations of the major sampling sites in South East Asia and Oceania are shown on each map; 1, Myanmar; 2, Thailand; 3, China; 4, Taiwan; 5, Aru Islands, Indonesia; and 6, Daru Island, Papua New Guinea. Wyndham and Darwin populations together form the Western Plateau. Populations where the studied LGT fragments are absent are in yellow, while those with the fragments are in black. The dashed black line indicates the approximate location of Wallace Line. The portion of the phylogenetic tree of each fragment corresponding to *A. semialata* is shown next to each map. Bayesian support values are indicated near nodes (* = 1.0), and the names of the individuals are indicated next to the location. See also Figures S1 and S2.

followed by a single successful dispersal to the east of the Wallace Line (black dashed line in Figure 1). Molecular dating indicates that this successful dispersal happened around 426-450 Ka (Figures 1 and S1), likely via land bridges that connected many Southeast Asia islands during glacial maxima [21, 22].

Population genetic analyses were focused on Australia, to track the spread of LGT within a single land mass. A total of 190 individuals from four regions from Australia (Wyndham, Darwin, Cairns, and Brisbane; Figure 1) and three Asian outgroups were sequenced with RAD markers. Based on these nuclear markers, each of the four Australian regions formed a monophyletic clade in the phylogeny (Figure S2A), where the two regions from the Western Plateau (Wyndham and Darwin) were sister, as were the two from the east of Australia (Cairns and Brisbane; Figure S2A). Multivariate and cluster analyses identified three genetic groups within Australia corresponding to the Western Plateau, Cairns, and Brisbane (Figures S2B and S2C; "populations" hereafter). The low-to-moderate genetic diversity within (mean $\pi$: 0.0007–0.0014; Figure 2) and divergence between (mean $F_{ST}$: 0.037–0.046; Figure 2) these three populations are suggestive of a genetic bottleneck during the initial colonization of Australia, followed by a relatively recent dispersal across the land mass (Figure 1).

## Laterally Acquired Fragments Are under Similar Selection Pressures as the Rest of the Recipient Genome

Mapping of reads onto the reference genome detected the presence of the three previously identified multigene LGT fragments [11] in all of the new whole-genome-sequenced accessions from Australia and nearby islands (i.e., locations 5 and 6 in Figure 1), and these were absent from Asian accessions other than the Philippines (Figure S3). Within the LGT fragments, the genetic diversity of both non-synonymous ($\pi_N$) and synonymous ($\pi_S$) sites was reduced compared with that of the rest of the genome ($\pi_{N,\ LGT}$ = 0.001 and $\pi_{S,\ LGT}$ = 0.002 versus $\pi_{N,\ WG}$ = 0.002 and $\pi_{S,\ WG}$ = 0.005), but the ratio ($\pi_N/\pi_S$) is similar between the two partitions (0.410 for LGT and 0.468 for the rest of the genome). These results show that LGT are fully integrated in the recipient genomes and behave as native DNA in terms of selection [7, 23].

## Successive Spread of LGT across Australia

Using sequences extracted from the genome datasets, phylogenetic trees were inferred for Oceanian (including Philippines) *A. semialata* and relatives of the putative grass donor species for the three LGT fragments. The relationships among *A. semialata* inferred from LGT_A are similar to those inferred from the organelles and genome-wide nuclear markers (Figures 1, S1, and S2A). The amount of mutations accumulated after the transfer of LGT_A, as estimated by the sum of branch lengths within *A. semialata*, is only slightly lower than for the rest of the genome, which might be linked to the initial LGT bottleneck (Figure 2). Population genomics show that despite this lower diversity, the differentiation and divergence among populations on the genomic region containing fragment LGT_A are not markedly different from the rest of the genome, and population-level analyses show an accumulation of mutations within *A. semialata* (Figures 2 and S4A). We conclude that fragment

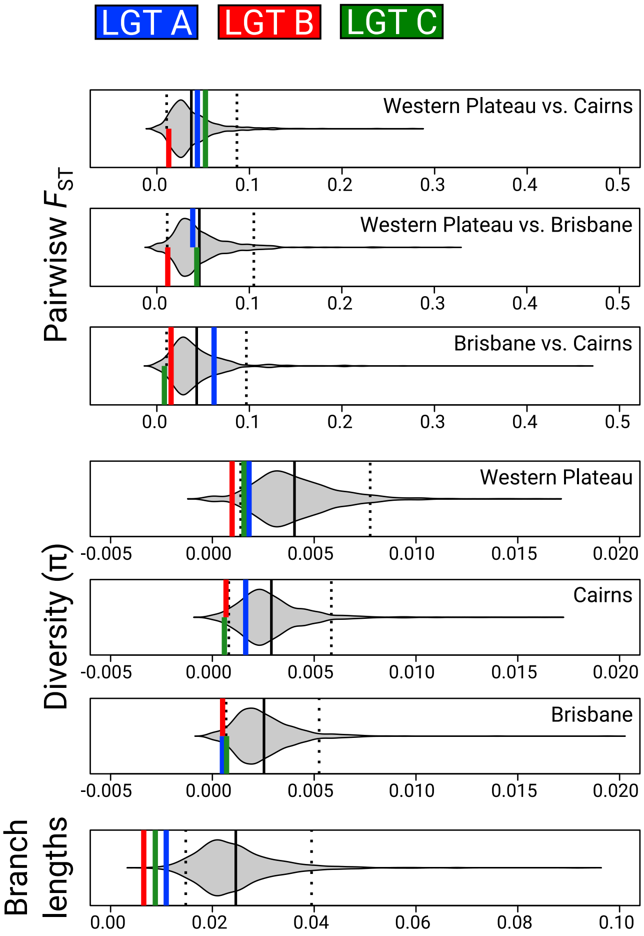


**Figure 2. Comparisons of Genomic Patterns between LGT and the Rest of the Genome.**

For three metrics ($F_{ST}$ between pairs of Australian populations, diversity within each population, and sum of phylogenetic branch lengths in Oceanian *Alloteropsis semialata*), the values for each of the three LGT fragments (blue, LGT_A; red, LGT_B; green, LGT_C) are compared to the rest of the genome. The bean plots show the distribution of the same metrics for 723 genomic fragments of similar size, with the 95% interval delimited with dashed black lines. See also Figure S4 and Table S3.

LGT_A was acquired early during the colonization of Oceania, around 426–450 ka, and then evolved alongside the rest of the recipient *A. semialata* genome (Figure 1). The integration of this fragment was likely mediated by strong selection for the gene that encodes a key enzyme of the photosynthetic pathway used by Australian and Filipino individuals of *A. semialata* (gene 1 in LGT_A; Figure 3) [5, 11, 24].

The relationships among *A. semialata* accessions inferred from fragment LGT_B differ from those based on organelles and genome-wide nuclear markers (Figures 1, S1, and S2A). The sum of branch lengths within *A. semialata* is very small for this fragment (Figure 2). Population genomics indicate that fragment LGT_B lies in a region of extremely low diversity and low differentiation between all pairs of populations (Figure 2), characterized by very weak differences in allele frequencies (Figure S4A). These patterns indicate a recent and rapid spread of LGT_B among established populations of *A. semialata* in Australia in the last 241 ka (Figure 1), as expected under the action of positive selection. This fragment includes one gene that is expressed at higher levels than the native orthologs (gene 4; Figure 3) [11], and this gene was likely the target of positive selection.

The phylogenetic and genomic patterns around LGT_C share some properties with each of the other two fragments, and this fragment also includes a gene expressed at higher level than the native ortholog (gene 9; Figure 3). The phylogenetic tree of LGT_C differs from those based on organelles and other nuclear markers, with one gene from Brisbane and the Philippines nested within accessions from Cairns (Figure 1). The sum of branch lengths is intermediate between the two other fragments, and interpopulation differentiation and allele divergence is very low for the Cairns-Brisbane comparison but more similar to the rest of the genome in the two other comparisons (Figures 2 and S4A). These patterns for fragment LGT_C suggest an older divergence among the Western Plateau and the east of Australia (Cairns and Brisbane) that probably matches the early diversification within Australia around 426 ka, but a very recent spread within the east of Australia and the Philippines (Figure 1).

## Genomic Erosion of LGT Fragments Reveals Hitchhiking of Neutral Genes

The mapping of whole-genome datasets allowed the composition of each LGT fragment to be compared with the reference genome. The presence-absence of multiple physically linked RAD markers within each fragment was then used to extrapolate the distribution of the structural variants detected in whole genomes to the 190 population-level samples. The conclusions based on RAD markers were subsequently verified using PCR for a subset of individuals, confirming that these markers can reliably identify large genomic variants. Fragment LGT_A was retrieved in full from all accessions (Figures 3 and S3; Data S1). In contrast, the length of fragments LGT_B and LGT_C varied between the reference genome and the other Australian accessions (Figures 3 and S3; Data S1), suggesting that genomic erosions occurred after the initial acquisitions. Eroded and complete variants were detected in all three Australian populations (Figure S3; Data S1), yet the eroded variants did not consistently group together in the phylogenetic trees (Figure 1). This suggests that the erosions appeared independently in the different populations or that recombination between the eroded and complete variants occurred within each of the populations.

Because there are protein-coding genes annotated to the eroded DNA segments of both LGT_B and LGT_C, the content of laterally acquired genes varies between individuals of *A. semialata* (Figure 3). From fragment LGT_B, three genes are present in all individuals, including the one expressed at higher levels than the native ortholog (gene 4 in LGT_B; Figure 3). The genetic differentiation between the two length variants of LGT_B is low within all Australian populations (Figure S4B), indicating that the variants are selectively equivalent despite their different gene contents. The three genes from the eroded fragment contain indels that disrupt the reading frames, indicating pseudogenization (Figure 3). These silencing mutations could have been selected for in parallel to gene losses, making the length variants of the LGT_B fragment selectively equivalent.

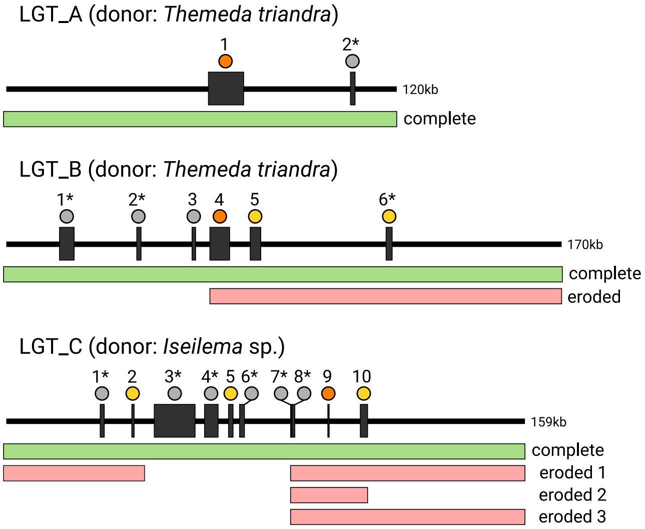


**Figure 3. Gene Composition of the Three LGT Fragments**

In each case, the black horizontal lines indicate the extent of the LGT fragment, and black boxes indicate the locations of annotated genes [10]. Circles above indicate the expression status in AUS1 [10]; gray, no expression; yellow, expression to a similar level as the native ortholog; orange, expression higher than the native equivalent. Genes within each fragment are numbered consecutively, and asterisks next to the numbers indicate pesudogenes. Horizontal boxes underneath each LGT fragment show the approximate extent of the length variants. See also Figure S3 and Data S1 and Table S1.

Only four of the ten genes from fragment LGT_C are present in all sequenced accessions, including the one expressed at higher levels than the native ortholog (gene 9 in LGT_C; Figure 3). Low diversity, no or low genetic differentiation, and no divergence were observed between the length variants of fragment LGT_C within the Western Plateau and Cairns populations, showing that the two variants are neutral relative to each other in these regions despite different gene contents (Figures 4 and S4C). The laterally acquired genes from the eroded segment are therefore neutral in *A. semialata*, yet some of them appear functional and are expressed at low levels (genes 2 and 5 in LGT_C; Figure 3) [11]. This demonstrates that neutral LGT are not immediately pseudogeneized, enabling their spread alongside the targets of positive selection, via genomic hitchhiking. In the cases reported here, the hitchhiked genes are also laterally acquired, but chance insertion of foreign DNA next to a selectively advantageous native allele would lead to similar genomic patterns and spread of an LGT fragment. We therefore conclude that genetic hitchhiking contributes to the integration of LGT in the recipient species, potentially contributing to their recurrence among grasses.

## Older Functional LGT Are Lost in Some Populations

Besides the three fragments restricted to Oceania and the Philippines, the reference genome contains seven older multigene LGT fragments that are shared with accessions from other geographic origins (Table S1) [11]. Erosion patterns have been identified for most of these fragments based on the comparison of geographically distant individuals [11], but the new genomic datasets for Australian populations revealed that five of the seven fragments also harbor gene presence-absence polymorphisms among these recently diverged populations, including for expressed genes (Table S1) [11]. This demonstrates that the genomic erosion detected on the three recent fragments is an ongoing process and continues long after the initial transfers. The origin of non-coding DNA separating the laterally acquired protein-coding genes is ambiguous in the older fragments because of subsequent recombination [11]. Focusing on the only fragment where enough RAD markers were present in protein-coding genes, a presence-absence polymorphism for a gene that is expressed was identified in all populations (gene 1 from LGT_N; Table S1). The intrapopulation genetic differentiation between the two length variants is low (Figure S4D), indicating that they are selectively equivalent in the Australian populations. These analyses of older LGT fragments indicate that hitchhiking allows the long-term persistence of functional LGT, some of which might have been originally adaptive and become neutral after the spread of *A. semialata* across Australia.

## Initially Neutral LGT Involved in Secondary Sweep

In stark contrast with populations from Cairns and the Western Plateau, there is a substantial increase in $F_{ST}$ between the samples with the complete versus eroded fragment LGT_C within Brisbane (Figure 4). This differentiation spans a 13.8 Mb region (approximately 3.8 Mb upstream and 9.8 Mb downstream of LGT_C) and is created by increased divergence ($d_{xy}$) in addition to a low diversity ($\pi$) (Figures S4C and S4E). The same genomic region shows a marked decrease in $F_{ST}$ between complete variants from Brisbane and Cairns (Figure 4), and the phylogenetic tree strongly suggests that the complete variant of fragment LGT_C has been introgressed from Cairns to Brisbane, in addition to the Philippines (Figure 1). In combination, population genomics and phylogenetic patterns indicate that the complete variant was recently introgressed from Cairns to Brisbane without subsequent recombination with the eroded variant in Brisbane. The high differentiation was likely facilitated by the location of LGT_C fragment to the putative centromere (Figure 4), a region with reduced recombination that accentuates population divergence [25–28].

The region that, in Brisbane and Cairns, is restricted to the complete variant of LGT_C includes the *SUMO E3 ligase SIZ2* gene (gene 2 in LGT_C; Figure 3), which has been implicated in the adaptive response to plant phosphate starvation [29]. We postulate that low soil phosphate levels in some localities around Brisbane and the Philippines have strongly selected for this gene, while environmental differences or mutations on other genes might decrease the adaptive value of the laterally acquired *SUMO E3 ligase SIZ2* gene in other regions of Australia. Indeed, genomic outliers between the Australian populations (Table S2) include several genes implicated in plant phosphate starvation so that preceding local adaptation might make selection on the acquired DNA context dependent.

Our analyses demonstrate that two functional genes from LGT_C were initially neutral and spread by genetic hitchhiking following positive selection on linked genes (genes 2 and 5; Figure 3). These initially neutral genes were then involved in a strong secondary sweep in Brisbane. The foreign DNA acquired from other plants, once fully integrated into the recipient genome, undergoes structural modification and accumulates mutations

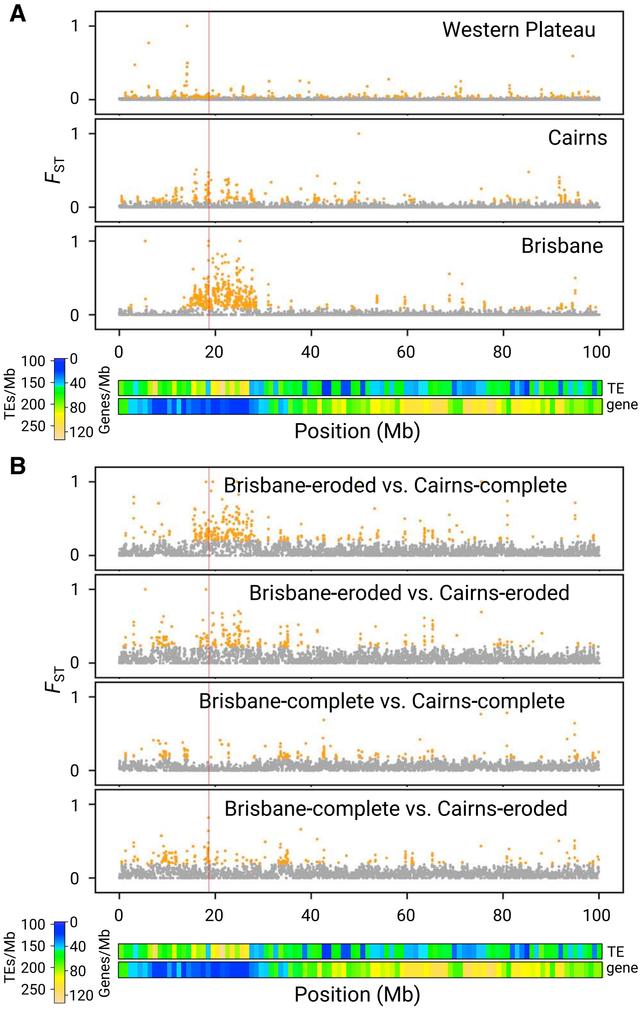


**Figure 4. Patterns of Differentiation among Length Variants of Fragment LGT_C**

The pairwise $F_{ST}$ is shown for 100 kilo-bases (kb) sliding windows (10 kb slides) along chromosome 9 between length variants of LGT_C.

(A) The differentiation between complete versus eroded variants within each of the three Australian populations is shown.

(B) The differentiation is shown for each variant between Cairns and Brisbane. Windows with $F_{ST}$ values above the 95th percentile are in orange. The cartoon underneath each figure shows the density of transposable elements (TE; top) and genes (bottom) per megabase (Mb). The vertical red lines indicate the position of LGT_C.

See also Figure S4.

as evidenced in the different LGT fragments we analyzed (Figures 1, 3, S3, and S4) and other study systems [6, 7, 23]. It therefore becomes part of the standing genetic variation in the recipient species, and it can subsequently be used for local adaptation at a later point in time, either by co-opting functional genes, as might be the case of *SUMO E3 ligase SIZ2*, or potentially by recruiting regulatory motifs and protein domains to generate new functions. We conclude that interspecific genetic exchanges among plants have long-lasting evolutionary consequences that go beyond the immediate acquisition of adaptive variants.

## Conclusions

Combining comparative genomics and population-level analyses, we show that plant-to-plant LGT fragments containing genes that added novel functions to the recipient were rapidly integrated in the recipient genome and spread among established populations. Laterally acquired fragments were then gradually eroded, and the resulting coexistence of length variants allowed us to narrow down which genes within the larger genomic fragments had been selected for. This revealed that laterally acquired genes that are selectively neutral can remain functional and spread by hitchhiking, being transmitted alongside positively selected genes that are physically linked. Therefore, genetic hitchhiking might contribute to the recurrence of successful LGT among some grasses. Importantly, even if the spread of some LGT is not initially driven by selection, laterally acquired genes that persist as neutral presence-absence polymorphisms can later contribute to local adaptation. In addition to providing novel functions, as widely reported [2, 3, 5–7, 11, 12, 14], LGT can therefore contribute to the adaptation of grasses by increasing their genomic diversity, which can later be used for the colonization of new habitats.

## STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- LEAD CONTACT AND MATERIALS AVAILABILITY
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - ○ Whole-genome sequencing and assembly of organelle genome
  - ○ Assembly of laterally acquired genomic fragments
  - ○ Population-level RAD sequencing
  - ○ Identification of structural variants of LGT fragments
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ○ Analyses of organelle genomes
  - ○ Phylogenetic analyses of laterally acquired genomic fragments
  - ○ Comparison of synonymous and non-synonymous sites
  - ○ Nuclear phylogenetic analyses based on RAD markers
  - ○ Overall population structure within Australia
  - ○ Patterns of differentiation across the genome and among structural variants of LGT fragments
- DATA AND SOFTWARE AVAILABILITY

## SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cub.2019.09.023.

## ACKNOWLEDGMENTS

This work was funded by NERC (grant NE/M00208X/1) and by the ERC (grant ERC-2014-STG-638333). P.A.C. and P.N. were supported by Royal Society University Research Fellowships (numbers URF120119 and URF138573, respectively). We thank Tricia Handasyde for collection of plant material from Wyndham, Australia, and Elsie Place for help with genome sequencing. Library preparation and sequencing were carried out by Sheffield Diagnostic Genetics Service, Sheffield Children's NHS Trust, Sheffield and Edinburgh Genomics, The University of Edinburgh. Edinburgh Genomics is partly supported through core grants from the NERC (R8/H10/56), MRC (MR/K001744/1), and BBSRC (BB/J004243/1).

## AUTHOR CONTRIBUTIONS

J.K.O., C.P.O., P.N., and P.A.C. designed the study. L.T.D., M.R.L., J.T., N.C., M.A., D.Y., C.P.O., and P.A.C. carried out field work. J.K.O. generated and analyzed the data with the help of L.T.D., H.J.B., G.S., and K.Z. J.K.O. and P.A.C. wrote the paper with the help of all authors.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 11, 2019
Revised: September 6, 2019
Accepted: September 10, 2019
Published: October 31, 2019

## REFERENCES

1. Davis, C.C., and Wurdack, K.J. (2004). Host-to-parasite gene transfer in flowering plants: phylogenetic evidence from Malpighiales. Science *305*, 676–678.
2. Richards, T.A., Soanes, D.M., Foster, P.G., Leonard, G., Thornton, C.R., and Talbot, N.J. (2009). Phylogenomic analysis demonstrates a pattern of rare and ancient horizontal gene transfer between plants and fungi. Plant Cell *21*, 1897–1911.
3. Moran, N.A., and Jarvik, T. (2010). Lateral transfer of genes from fungi underlies carotenoid production in aphids. Science *328*, 624–627.
4. Yoshida, S., Maruyama, S., Nozaki, H., and Shirasu, K. (2010). Horizontal gene transfer by the parasitic plant Striga hermonthica. Science *328*, 1128.
5. Christin, P.A., Edwards, E.J., Besnard, G., Boxall, S.F., Gregory, R., Kellogg, E.A., Hartwell, J., and Osborne, C.P. (2012). Adaptive evolution of C($_4$) photosynthesis through recurrent lateral gene transfer. Curr. Biol. *22*, 445–449.
6. Li, F.W., Villarreal, J.C., Kelly, S., Rothfels, C.J., Melkonian, M., Frangedakis, E., Ruhsam, M., Sigel, E.M., Der, J.P., Pittermann, J., et al. (2014). Horizontal transfer of an adaptive chimeric photoreceptor from bryophytes to ferns. Proc. Natl. Acad. Sci. USA *111*, 6672–6677.
7. Yang, Z., Zhang, Y., Wafula, E.K., Honaas, L.A., Ralph, P.E., Jones, S., Clarke, C.R., Liu, S., Su, C., Zhang, H., et al. (2016). Horizontal gene transfer is more frequent with increased heterotrophy and contributes to parasite adaptation. Proc. Natl. Acad. Sci. USA *113*, E7010–E7019.
8. Mahelka, V., Krak, K., Kopecký, D., Fehrer, J., Šafář, J., Bartoš, J., Hobza, R., Blavet, N., and Blattner, F.R. (2017). Multiple horizontal transfers of nuclear ribosomal genes between phylogenetically distinct grass lineages. Proc. Natl. Acad. Sci. USA *114*, 1726–1731.
9. Peccoud, J., Loiseau, V., Cordaux, R., and Gilbert, C. (2017). Massive horizontal transfer of transposable elements in insects. Proc. Natl. Acad. Sci. USA *114*, 4721–4726.
10. Vogel, A., Schwacke, R., Denton, A.K., Usadel, B., Hollmann, J., Fischer, K., Bolger, A., Schmidt, M.H.W., Bolger, M.E., Gundlach, H., et al. (2018). Footprints of parasitism in the genome of the parasitic flowering plant *Cuscuta campestris*. Nat. Commun. *9*, 2515.
11. Dunning, L.T., Olofsson, J.K., Parisod, C., Choudhury, R.R., Moreno-Villena, J.J., Yang, Y., Dionora, J., Quick, W.P., Park, M., Bennetzen, J.L., et al. (2019). Lateral transfers of large DNA fragments spread functional genes among grasses. Proc. Natl. Acad. Sci. USA *116*, 4416–4425.
12. Milner, D.S., Attah, V., Cook, E., Maguire, F., Savory, F.R., Morrison, M., Müller, C.A., Foster, P.G., Talbot, N.J., Leonard, G., and Richards, T.A. (2019). Environment-dependent fitness gains can be driven by horizontal gene transfer of transporter-encoding genes. Proc. Natl. Acad. Sci. USA *116*, 5613–5622.
13. Reiss, D., Mialdea, G., Miele, V., de Vienne, D.M., Peccoud, J., Gilbert, C., Duret, L., and Charlat, S. (2019). Global survey of mobile DNA horizontal transfer in arthropods reveals Lepidoptera as a prime hotspot. PLoS Genet. *15*, e1007965.
14. Zhang, Q., Chen, X., Xu, C., Zhao, H., Zhang, X., Zeng, G., Qian, Y., Liu, R., Guo, N., Mi, W., et al. (2019). Horizontal gene transfer allowed the emergence of broad host range entomopathogens. Proc. Natl. Acad. Sci. USA *116*, 7982–7989.
15. Prentice, H.C., Li, Y., Lönn, M., Tunlid, A., and Ghatnekar, L. (2015). A horizontally transferred nuclear gene is associated with microhabitat variation in a natural plant population. Proc. Biol. Sci. *282*, 20152453.
16. Olofsson, J.K., Bianconi, M., Besnard, G., Dunning, L.T., Lundgren, M.R., Holota, H., Vorontsova, M.S., Hidalgo, O., Leitch, I.J., Nosil, P., et al. (2016). Genome biogeography reveals the intraspecific spread of adaptive mutations for a complex trait. Mol. Ecol. *25*, 6107–6123.
17. Roulin, A., Piegu, B., Wing, R.A., and Panaud, O. (2008). Evidence of multiple horizontal transfers of the long terminal repeat retrotransposon RIRE1 within the genus *Oryza*. Plant J. *53*, 950–959.
18. Vallenback, P., Ghatnekar, L., and Bengtsson, B.O. (2010). Structure of the natural transgene *PgiC2* in the common grass *Festuca ovina*. PLoS ONE *5*, e13529.
19. El Baidouri, M., Carpentier, M.C., Cooke, R., Gao, D., Lasserre, E., Llauro, C., Mirouze, M., Picault, N., Jackson, S.A., and Panaud, O. (2014). Widespread and frequent horizontal transfers of transposable elements in plants. Genome Res. *24*, 831.838.
20. Lundgren, M.R., Besnard, G., Ripley, B.S., Lehmann, C.E., Chatelet, D.S., Kynast, R.G., Namaganda, M., Vorontsova, M.S., Hall, R.C., Elia, J., et al. (2015). Photosynthetic innovation broadens the niche within a single species. Ecol. Lett. *18*, 1021–1029.
21. Pillans, B., Chappell, J., and Naish, T.R. (1998). A review of the Milankovitch climatic beat: template for Plio-Pleistocene sea-level changes and sequence stratigraphy. Sediment. Geol. *122*, 5–21.
22. Voris, H.K. (2000). Maps of Pleistocene sea levels in Southeast Asia: shorelines, river systems and time durations. J. Biogeogr. *27*, 1153–1167.
23. Zhang, Y., Fernandez-Aparicio, M., Wafula, E.K., Das, M., Jiao, Y., Wickett, N.J., Honaas, L.A., Ralph, P.E., Wojciechowski, M.F., Timko, M.P., et al. (2013). Evolution of a horizontally acquired legume gene, albumin 1, in the parasitic plant *Phelipanche aegyptiaca* and related species. BMC Evol. Biol. *13*, 48.
24. Dunning, L.T., Lundgren, M.R., Moreno-Villena, J.J., Namaganda, M., Edwards, E.J., Nosil, P., Osborne, C.P., and Christin, P.A. (2017). Introgression and repeated co-option facilitated the recurrent emergence of C$_4$ photosynthesis among close relatives. Evolution *71*, 1541–1555.
25. Charlesworth, B., Morgan, M.T., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. Genetics *134*, 1289–1303.
26. Charlesworth, B., Nordborg, M., and Charlesworth, D. (1997). The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genet. Res. *70*, 155–174.
27. Round, E.K., Flowers, S.K., and Richards, E.J. (1997). *Arabidopsis thaliana* centromere regions: genetic map positions and repetitive DNA structure. Genome Res. *7*, 1045–1053.
28. Berner, D., and Roesti, M. (2017). Genomics of adaptive divergence with chromosome-scale heterogeneity in crossover rate. Mol. Ecol. *26*, 6351–6369.
29. Miura, K., Rus, A., Sharkhuu, A., Yokoi, S., Karthikeyan, A.S., Raghothama, K.G., Baek, D., Koo, Y.D., Jin, J.B., Bressan, R.A., et al. (2005). The Arabidopsis SUMO E3 ligase SIZ1 controls phosphate deficiency responses. Proc. Natl. Acad. Sci. USA *102*, 7760–7765.

30. Catchen, J., Hohenlohe, P.A., Bassham, S., Amores, A., and Cresko, W.A. (2013). Stacks: an analysis tool set for population genomics. Mol. Ecol. *22*, 3124–3140.

31. Rambaut, A., Suchard, M.A., Xie, W., and Drummond, A.J. (2013). Tracer v1.6, Available from. http://tree.bio.ed.ac.uk/software/tracer/.

32. Patel, R.K., and Jain, M. (2012). NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS ONE *7*, e30619.

33. Schmieder, R., and Edwards, R. (2011). Quality control and preprocessing of metagenomic datasets. Bioinformatics *27*, 863–864.

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

35. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup (2009). The sequence alignment/map format and SAMtools. Bioinformatics *25*, 2078–2079.

36. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993.

37. Nosil, P., Gompert, Z., Farkas, T.E., Comeault, A.A., Feder, J.L., Buerkle, C.A., and Parchman, T.L. (2012). Genomic consequences of multiple speciation processes in a stick insect. Proc. Biol. Sci. *279*, 5058–5065.

38. Parchman, T.L., Gompert, Z., Mudge, J., Schilkey, F.D., Benkman, C.W., and Buerkle, C.A. (2012). Genome-wide association genetics of an adaptive trait in lodgepole pine. Mol. Ecol. *21*, 2991–3005.

39. Petersen, B.K., Weber, J.N., Kay, E.H., Fisher, H.S., and Hoekstra, H.E. (2014). Double digest RADseq: An inexpensive methode for *de novo* SNP discovery and genotyping in model and non-model species. PLoS ONE *7*, e37135.

40. Olofsson, J.K., Cantera, I., Van de Paer, C., Hong-Wa, C., Zedane, L., Dunning, L.T., Alberti, A., Christin, P.A., and Besnard, G. (2019). Phylogenomics using low-depth whole genome sequencing: a case study with the olive tribe. Mol. Ecol. Resour. *19*, 877–892.

41. Wood, D.P., Olofsson, J.K., McKenzie, S.W., and Dunning, L.T. (2018). Contrasting phylogeopgraphic structures between freshwater lycopods and angiosperms in the British Isles. Bot. Lett. *165*, 476–486.

42. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120.

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

44. R Core Team (2018). R: A language and environment for statistical computing. (R Foundation for Statistical Computing).

45. Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics *30*, 1312–1313.

46. Jombart, T. (2008). adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics *24*, 1403–1405.

47. Jombart, T., and Ahmed, I. (2011). adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics *27*, 3070–3071.

48. Skotte, L., Korneliussen, T.S., and Albrechtsen, A. (2013). Estimating individual admixture proportions from next generation sequencing data. Genetics *195*, 693–702.

49. Kopelman, N.M., Mayzel, J., Jakobsson, M., Rosenberg, N.A., and Mayrose, I. (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. Mol. Ecol. Resour. *15*, 1179–1191.

50. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al.; 1000 Genomes Project Analysis Group (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158.

51. Hunter, J.D. (2007). Mathplotlin: A 2D graphical environment. Comput. Sci. Eng. *9*, 90–95.

52. Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., and Huelsenbeck, J.P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. *61*, 539–542.

53. Drummond, A.J., and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. *7*, 214.

54. Evanno, G., Regnaut, S., and Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol. Ecol. *14*, 2611–2620.

55. Ma, T., Wang, K., Hu, Q., Xi, Z., Wan, D., Wang, Q., Feng, J., Jiang, D., Ahani, H., Abbott, R.J., et al. (2018). Ancient polymorphisms and divergence hitchhiking contribute to genomic islands of divergence within a poplar species complex. Proc. Natl. Acad. Sci. USA *115*, E236–E243.

56. Dunning, L.T., Moreno-Villena, J.J., Lundgren, M.R., Dionora, J., Salazar, P., Adams, C., Nyirenda, F., Olofsson, J.K., Mapaura, A., Grundy, I.M., et al. (2019). Key changes in gene expression identified for different stages of $C_4$ evolution in *Alloteropsis semialata*. J. Exp. Bot. *70*, 3255–3268.

## STAR★METHODS

### KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Biological Samples | | |
| Wild collected individuals of *Alloteropsis semialata* | See Table S4 | N/A |
| Chemicals, Peptides, and Recombinant Proteins | | |
| EcoRI (20,000 units/mL) | New England BioLabs | R0101L |
| MseI (10,000 units/mL) | New England BioLabs | R0525L |
| T4 DNA ligase (400,000 units/mL) | New England BioLabs | M0202L |
| Iproof High Fidelity DNA polymerase | BioRad | Cat no. 172-5301 |
| QIAquick gel extraction kit | QIAGEN | Cat no. 28706 |
| DNeasy Plant Mini Kit (250) | QIAGEN | Cat no. 69106 |
| NEBNex® UltraTM II DNA Library Prep Kit for Illumina® | New England BioLabs | E7103 |
| Deposited Data | | |
| Cleaned and demultiplexed sequencing reads | This paper | BioProject PRJNA560360 |
| *Alloteropsis semialata* reference genome (ASEM_AUS1_v1.0) | [11] | GenBank accession QPGU01000000 |
| Oligonucleotides | | |
| EcoRI adapters | [30] | https://datadryad.org/stash/dataset/doi:10.5061/dryad.m2271pf1 |
| MseI_P2.1<br>GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT | This paper, and [31] | N/A |
| MseI_P2.2<br>TAAGATCGGAAGAGCGAGAACAA | This paper, and [31] | N/A |
| PCR2_Idx_1_ATCACG (Illumina sequencing primer, index 1)<br>CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAGTTCAGACGTGTGC | | N/A |
| Illpcr1 (Illumine sequencing primer)<br>A*A*TGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT | | N/A |
| 11 pairs of PCR primers | See Table S1 | N/A |
| Software and Algorithms | | |
| NGS QC Toolkit v 2.3.3 | [32] | http://14.139.61.3:8080/ngsqctoolkit/ |
| Prins-Seq-lite v.0.20.3 | [33] | http://prinseq.sourceforge.net/ |
| Bowtie2 v.2.2.3 | [34] | http://bowtie-bio.sourceforge.net/bowtie2/index.shtml |
| MrBayes 2.3.2.6 | [35] | http://nbisweden.github.io/MrBayes/ |
| Tracer | [36] | http://tree.bio.ed.ac.uk/software/tracer/ |
| Beast 1.8.4 | [37] | https://beast.community/ |
| SAMtools v.2.2.3 | [38, 39] | http://www.htslib.org/ |
| published bash scripts | This paper and [16, 40] | https://github.com/jill-olofsson/low-depth-sequencing_analyses |
| APE | [41] | http://ape-package.ird.fr/ |
| R v.3.4.4 | [42] | https://www.R-project.org/. |
| BCFtools v.1.3.1 | [39] | https://samtools.github.io/bcftools/ |
| Trimmomatic tool kit | [43] | http://www.usadellab.org/cms/?page=trimmomatic |
| processRADtag.pl script from the program STACKS | [44] | http://catchenlab.life.illinois.edu/stacks/ |
| RAxML v.8.2.11 | [45] | https://cme.h-its.org/exelixis/web/software/raxml/ |

(*Continued on next page*)

***Continued***

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| adegenet | [46, 47] | http://adegenet.r-forge.r-project.org/ |
| NGSadmix | [48] | http://www.popgen.dk/software/index.php/NgsAdmix |
| CLUMPAK | [49] | http://clumpak.tau.ac.il/ |
| VCFtools v.0.1.14 | [50] | http://vcftools.sourceforge.net/ |
| matplotlib v 3.1.1 | [51] | https://matplotlib.org/ |

## LEAD CONTACT AND MATERIALS AVAILABILITY

Further information or materials associated with this research will be made available upon request to the lead contact author, Dr. Pascal-Antoine Christin (p.christin@sheffield.ac.uk). This study did not generate any new, or unique reagents.

## EXPERIMENTAL MODEL AND SUBJECT DETAILS

Four accessions of *Alloteropsis semialata* (R. Br.) Hitchc. used for whole-genome sequencing were obtained from herbaria to increase the coverage of mainland Asia [voucher nos. McKee 6138 (K) – Myanmar and Koelz 32989 (K) – India] and islands on the continental shelf north of Australia [voucher nos. Brass 6343 (K, BRI) – Daru Island, and von Balgooy & Mamesah 6270 (K) – Aru Islands]. All other samples were collected in the field. Individuals of *A. semialata* were collected in multiple localities from four regions of Australia; Wyndham, Darwin, Cairns, and Brisbane (Figure 1; Table S3). Places where *A. semialata* grew were identified based on herbarium records or field surveys. A similar strategy was used to obtain population-level samples from Sri Lanka, the Philippines, and Taiwan (Table S3). For each locality, GPS coordinates were recorded, and leaves from up to five distinct individuals were collected and dried in silica gel. DNA was extracted using the DNeasy Plant Mini kit (QIAGEN).

## METHOD DETAILS

### Whole-genome sequencing and assembly of organelle genome

The whole genomes of 11 samples of *Alloteropsis semialata* were sequenced in this work (Figure S1). Libraries for the four samples retrieved from herbaria were built and sequenced at the Genopole platform in Toulouse (France), with low target coverage (< 5x) and 150 bp paired-end Illumina reads [16]. Because of the fragmented nature of DNA recovered from herbarium samples, the insert sizes were relatively small (< 350bp). The other seven samples, available as silica gel dried leaves, included one sample from Sri Lanka and six from different regions of Australia (Figure S1). The Australian individuals were selected for genome sequencing because preliminary analyses of RAD-markers suggested they contained structural variants of laterally acquired genomic fragments as compared to the two Australian genomes previously characterized (AUS1 and AUS2-1) [11]. For these accessions, approximately 60-680 ng of DNA was used to build sequencing libraries with a target insert size of 550 bp using the NEBNex® UltraTM II DNA Library Prep Kit for Illumina® (New England Biolabs). Each library was paired-end (250 bp reads) sequenced on 1/6th of an Illumina® HiSeq2500 lane in the rapid output mode at Sheffield Diagnostic Genetics Service (Sheffield, UK) for a coverage of ~10x. Equivalent datasets were retrieved from previous work [11] for 32 other samples of *A. semialata* spread across Africa, Asia, and Oceania and ten other grass samples belonging to potential LGT donor species (samples from *Themeda*, *Iseilema* and *Heteropogon*).

Raw de-multiplexed reads were quality filtered and trimmed using NGS QC Toolkit v 2.3.3 [32] as previously described [11]. In short, sequencing adaptors, reads with < 80% high quality bases (Q > 20), and reads with ambiguous bases were removed and the remaining reads were trimmed removing low-quality bases (Q < 20) from the 3′ end. Finally, Prins-Seq-lite v.0.20.3 [33] was used to remove duplicated reads. Cleaned reads were then mapped to the reference genome of *A. semialata* (ASEM_AUS1_v1.0; GenBank accession QPGU01000000) [11] using Bowtie2 v.2.2.3 [34] with default parameters for single-end or paired-end reads, as appropriate. The maximum insert size was increased to 1,000 bp for the seven samples extracted from silica gel dried leaves to compensate for the larger insert size of these libraries.

Reads of each of the *A. semialata* samples mapping uniquely to one of the two organelle genomes from the reference genome were retrieved and consensus sequences were obtained, using previously published bash scripts [16, 40]. All heterozygous sites were converted to missing data to avoid interpreting positions affected by organelle-nuclear transfers. The alignment of each of the organelle genomes was trimmed, removing positions missing in > 10% of the samples.

### Assembly of laterally acquired genomic fragments

Base calls for the three genomic fragments laterally acquired recently (LGT fragments LGT_A, LGT_B, and LGT_C) were extracted from mapped reads using the mpileup function in SAMtools v.2.2.3 [35, 36] for all *A. semialata* samples where the fragments were

Please cite this article in press as: Olofsson et al., Population-Specific Selection on Standing Variation Generated by Lateral Gene Transfers in a Grass, Current Biology (2019), https://doi.org/10.1016/j.cub.2019.09.023

detected as well as potential donor species. Previously published bash scripts were then used to call consensus sequences from each LGT fragment, adjusting for differences in coverage [16, 40]. Heterozygous sites were coded as ambiguous nucleotides using the IUPAC codes. For high coverage data (> 40x), only positions with an overall depth of at least ten and where each allele had an individual depth of at least four were kept. For medium coverage data (5-25x), only positions with an overall depth of at least three and where each allele had an individual depth of at least two were kept. Finally, for low coverage data (< 5x) no depth filters were implemented as these would remove all positions [16]. Phylogenetic trees were inferred from the alignments for each of the LGT fragment as described above for the organelle genomes.

### Population-level RAD sequencing

Double digested restriction associated DNA libraries were produced following existing protocols [37, 38], with some alterations. The common adaptor was modified to allow for paired-end sequencing and the number of PCR cycles was reduced to 16 [39, 41]. In short, a total of 7 μl of DNA extraction (approximately 200-700 ng DNA) was digested with EcoRI and *Mse*I at 37°C for 8 h. Barcoded adaptors were then ligated to the EcoRI overhang and a common adaptor was ligated to the *Mse*I overhang at 16°C for 6 h. Ligation products were finally PCR amplified (16 cycles) using standard Illumina® sequencing primers. A total of 93-96 barcode compatible libraries from the same or different projects were pooled in equal volumes and the pools were size selected (300-600 bp) by gel extraction using the QIAquick Gel Extraction Kit (QIAGEN) following the manufacturer's protocol. Each size selected library pool was paired-end sequenced (125 bp) on one lane of an Illumina® HiSeq2500 at Edinburgh Genomics, University of Edinburgh (UK), following standard protocols.

Adaptors and primers were removed from the raw sequencing reads using the ILLUMINACLIP option in palindrome mode in the Trimmomatic tool kit [42], supplying the program with the known adaptor and primer sequences. Trimmomatic was further used to remove low quality bases (Q < 3) from the 5′ and 3′ ends, as well as all bases with a quality score below 15 in a four-base sliding window. The cleaned reads were de-muliplexed and barcodes were removed using the processRADtag.pl script from the program STACKS [30].

Cleaned reads were mapped to the reference genome using the default parameters for paired-end reads in Bowtie2. Variants were called from uniquely mapped reads using the mpileup function in SAMtools and the consensus calling function in BCFtools. All variants were merged in SAMtools and filtered to only keep positions with less than 90% missing data in each of the three population clusters in Australia and each of the Philippines, Sri Lanka and Taiwan populations.

### Identification of structural variants of LGT fragments

The coverages from each whole-genome dataset were plotted on each of the three recent LGT fragments and 0.1 Mb downsteam and upstream of putative break points between LGT and native DNA as previously determined [11]. The results were used to establish the presence/absence of each LGT fragment in each individual, and to detect length variants among the whole-genome sequenced samples. The reference genome of the Australian individual from *A. semialata* (AUS1) contains other multigene LGT that are shared with accessions from other geographic regions and are therefore older [11]. Because of longer divergence times and recombination with native DNA, non-coding DNA on these older fragments cannot be confidently assigned to native DNA versus LGT [11]. For these older fragments, we consequently only used the mapping to establish the presence/absence of each protein-coding gene among newly whole-genome sequenced samples, following the method of Dunning et al. [11].

The RAD markers were subsequently used to extrapolate the observed erosion patterns to the population-level samples. The presence/absence of RAD markers mapping to the previously defined LGT genomic boundaries was recorded. Only RAD markers with a length above 100 bp and with less than 90% missing data among samples were considered. In total, these 79,813 markers cover 13,556,475 bp (1.81% of the assembled genome). A stringent approach was adopted, drawing conclusions only from the joint presence/absence of a minimum of three physically linked markers, and subsequently verifying the patterns with a PCR approach (see below).

For fragment LGT_A there was a total of one intergenic and 10 intragenic RAD markers, with all intragenic markers located in the gene ASEM_AUS1_12633 (Data S1). No length variant was detected for this fragment. All samples with at least three intragenic RAD markers were estimated to have the complete LGT_A fragment. All other samples were considered ambiguous (Data S1). One sample from the Western Plateau fell in this ambiguous category.

For fragment LGT_B only one RAD marker was intragenic (located in gene ASEM_AUS1_26621 [gene 1 in Figure 3]; Data S1). This marker was positioned in the eroded fragment inferred from whole-genome sequencing data, as were another six RAD markers positioned in intergenic regions between ASEM_AUS1_26621 and ASEM_AUS1_07323 (gene 4 in Figure 3; Data S1). Samples with none of these seven RAD markers were estimated to have an eroded fragment LGT_B and hence to be missing the three LGT genes located in this segment (Figure 3; Data S1). Samples with at least three of these seven markers were estimated to have the complete fragment LGT_B. The erosion status of samples not fulfilling any of these two criteria was considered ambiguous and these samples were not included in analyses comparing length variants of fragment LGT_B. A total of nine samples from the Western Plateau, four samples from Cairns, and six samples from Brisbane fell in this ambiguous category.

For fragment LGT_C, a total of five RAD markers were intragenic, four within ASEM_AUS1_20550 (gene 3 in Figure 3) and one within ASEM_AUS1_25959A/B (genes 7 and 8 in Figure 3; Data S1). Samples with no RAD markers before the location of ASEM_AUS1_20547 (gene 6 in Figure 3), out of the 16 that were expected, were estimated to miss six genes from fragment LGT_C (eroded 2 or eroded 3 in Figure 3; Data S1). Samples with at least two RAD-markers on one side and at least one on the other side of ASEM_AUS1_20533 (gene 1 in Figure 3) before ASEM_AUS1_20551 (gene 2 in Figure 3), but no intragenic RAD marker within ASEM_AUS1_20550 (gene 3 in Figure 3), and no intergenic marker between ASEM_AUS1_20550 (gene 3 in Figure 3) and ASEM_AUS1_20548 (gene 5 in Figure 3) were determined to miss four genes from fragment LGT_C (Figure 3; Data S1). Samples with at least two RAD markers on one side and at least one on the other side of ASEM_AUS1_20533 (gene 1 in Figure 3), and at least three markers over the region spanning the coding gene ASEM_AUS1_20550 (gene 3 in Figure 3) and the non-coding DNA before ASEM_AUS1_20548 (gene 5 in Figure 3) were determined to have a complete LGT_C fragment (Figure 3; Data S1). The erosion status of all other samples was considered ambiguous and these samples were not included in analyses comparing length variants of fragment LGT_C. A total of 17 samples from the Western Plateau, 18 samples from Cairns, and 14 samples from Brisbane fell in this ambiguous category.

RAD markers were similarly reported for seven older fragments, but only intragenic RAD markers were considered when extrapolating the patterns observed on whole genomes because the origin of intergenic markers is unknown for these older LGT [11]. No RAD marker was available for fragment LGT_U. Fragments LGT_J and LGT_S had no intragenic RAD marker, LGT_G and LGT_M had only one intragenic RAD marker, while the four intragenic RAD markers from LGT_V were in physically distant genes (Data S1). We consequently only extrapolated the whole-genome patterns to the population-level samples for LGT_N. In this fragment, three RAD markers were located within ASEM_AUS1_28575, a gene that is present in only some individuals according to the whole-genome analyses (gene 1 in Table S1). The gene was considered present if all three markers were detected and absent if none was present. All other samples were considered as ambiguous. Because of the low number of RAD markers, two samples from the Philippines, one from Taiwan, 43 from the Western Plateau, 23 from Cairns, and 41 from Brisbane fell in this category.

The erosion patterns deduced from the RAD markers were verified using a PCR approach for a sample of 30 individuals per fragment, selected to capture ten individuals from each of the three Australian populations with a diversity of erosion variants. Pairs of primers, where possible spanning the coding/non-coding boundaries, were designed based on the reference genome for each gene from the two recent LGT fragments with variation within some populations (LGT_B, and LGT_C). For each of the primer pairs, a PCR was conducted on the 30 selected samples plus a positive control (AUS1; reference genome), a negative control (RSA3-1 sample from South Africa without the LGT [11]) and a blank. Amplification was done using the QIAGEN Multiplex PCR Kit, with ~5-10 ng of DNA, 0.4 $\mu$L of each primer (10 $\mu$M) and 10 $\mu$L of 2x Master Mix in a final volume of 20 $\mu$l. PCR conditions consisted of an initial denaturation at 95°C for 15 min followed by 35 cycles of denaturation at 94°C for 30 s, annealing at 63°C for 30-60 s and extension at 72°C for 30-60 s, with a final extension step at 72°C for 10 min (Data S1). Primers allowed the amplification of five genes from LGT_B (genes 1 and 2 in the eroded segment and genes 4-6 in the non-eroded segment; Figure 3) and six genes from LGT_C (genes 1-4 in the eroded segment and genes 9 and 10 in the non-eroded segment; Figure 3). In each case, positive PCR products detected on agarose gels and matching the expected fragment size were recorded. None of the 84 genes predicted to be absent based on RAD markers amplified, while only 11 of the 246 genes predicted to be present failed to amplify. These PCR failures can stem from additional erosion patterns or mutations in the primer-binding sites, but only one of the 60 individuals (30 for each fragment) would be assigned to a different genomic variant based on the PCR results.

## QUANTIFICATION AND STATISTICAL ANALYSIS

### Analyses of organelle genomes

The trimmed alignments were used to infer Bayesian phylogenetic trees in MrBayes 2.3.2.6 [52], with a GTR+G substitution model. Analyses were run for 10,000,000 generations, sampling a tree every 1,000 generations after a burn-in period of 2,500,000. Convergence of the runs and adequacy of the burn-in period were verified using Tracer [31]. Majority-rule consensus trees were obtained from the posterior trees.

Molecular dating was performed on the chloroplast alignment using the Bayesian approach implement in Beast 1.8.4 [53]. The monophyly of each of the clades FG and ABCDE was enforced to root the tree (Figure S1) [16, 20]. The analysis was run for 100,000,000 generations, sampling a tree every 10,000 generations. Tracer was used to verify the convergence of the runs, check the adequacy of the burn-in period set to 10,000,000 generations, and monitor the ages for nodes representing the migration of Australia (Figures 1 and S1). The analysis was repeated with a strict clock and a log-normal relaxed clock, in each case using a coalescent prior with either a constant size or an exponential growth. While the relaxed clock leads to wider confidence intervals, all models were congruent, and results of the strict clock with a constant size prior are discussed.

### Phylogenetic analyses of laterally acquired genomic fragments

For each of the three LGT fragments, the amount of substitutions accumulated by *A. semialata* was estimated by summing all branch lengths from the most recent common ancestor of the species to the tips, using the package APE [43] in R v.3.4.4 [44]. Similar sums of branches were calculated for native genomic fragments of the same size. For this, consensus sequences of 723 genomic windows of 150 kb (average size of the studied LGT fragments) separated by 1 Mb were obtained with the same method for the same samples, providing a genome-wide distribution of the amount of mutations accumulated.

### Comparison of synonymous and non-synonymous sites

Genotypes covering the whole genome were obtained for the eight medium or high coverage whole-genome samples from Australia using a combination of the mpileup function in SAMtools and the consensus calling function in BCFtools v.1.3.1 [36]. For all annotated genes in the reference genome, four-fold degenerate sites were identified and considered as synonymous sites, while zero-fold degenerate sites were considered as those where mutations are non-synonymous. The diversity was then calculated independently for the synonymous sites ($\pi_S$) and non-synonymous sites ($\pi_N$). These metrics were obtained for two genome partitions; 1) the three recent LGT fragments containing multiple genes and 2) the rest of the genome.

### Nuclear phylogenetic analyses based on RAD markers

Variants with an overall minor allele count above 20 (minor allele frequency of approximately 0.05) were extracted from RAD sequencing data and used to infer a maximum likelihood phylogeny in RAxML v.8.2.11 [45], using a GTR+G substitution model. Node support was evaluated with 100 fast bootstrap pseudoreplicates.

### Overall population structure within Australia

The population connectivity within the single landmass of Australia was evaluated using admixture and principal component analyses. These analyses considered a total of 20,091 SNPs, which were detected in Australia and were at least 1 kb apart on the chromosomes. These were obtained after setting the minor allele count threshold to 17, to keep the minor allele frequency to ~0.05 with this sampling excluding other Asian locations. The diversity was described using a principal component analysis as implemented in the R package adegenet [44, 46, 47]. The Bayesian approach implemented in NGSadmix [48] was then used to assign each individual to one of K genetic clusters, with K varying between two and ten. Ten analytical replicates were obtained for each K value. The optimal number of clusters K was identified using the Evanno et al. (2005) method [54] as implemented in CLUMPAK [49].

### Patterns of differentiation across the genome and among structural variants of LGT fragments

The PCA and clustering analyses identified three main populations within Australia, corresponding to the Western Plateau (Wyndham and Darwin), Cairns, and Brisbane. Population statistics were calculated in sliding windows along the genome to assess the genomic differentiation between pairs of these populations.

Pairwise $F_{ST}$ were calculated in 100 kb overlapping windows, with 10 kb sliding steps, using VCFtools v.0.1.14 [50] and plotted using the python library matplotlib [51]. Outlier windows were determined from null distributions of pairwise $F_{ST}$ adjusted for number of variants within each window, using a method analogous to Ma et al. [55]. In short, mean $F_{ST}$ were calculated from variants resampled 1,000 times without replacement using a combination of custom bash scripts and VCFtools. Windows with observed values above the 95th percentile of the null distributions were considered $F_{ST}$ outliers. The function of proteins encoded by genes positioned in outlier loci was reported, with the expression pattern extracted from [56].

Absolute divergence ($d_{xy}$) was calculated from allele frequency estimates in 10 kb non-overlapping windows using a combination of custom python, R, and bash scripts. In short, allele frequencies were estimated at all positions for each population cluster using VCFtools. Positions with more than 90% missing data within each population were excluded and $d_{xy}$ was calculated for each position and population comparison. The mean $d_{xy}$ was then calculated per window, and plotted along the genome using the python library matplotlib. Null distributions were obtained by repeating the estimates on 100 datasets consisting of reshuffled genomic positions using custom bash and R scripts. Windows above the 99th percentile of null distributions were considered outliers.

Genetic diversity ($\pi$) within each population cluster was calculated in overlapping 100 kb windows, with 10 kb sliding steps, using VCFtools. It was plotted along the genome using the python library matplotlib.

We further used differences in allele frequencies between the three populations as a proxy for the amount of allele sharing among pairs of populations. Allele frequencies were computed within each population in VCFtools as outlined above. The differences in allele frequencies between populations were then calculated and plotted along the genome using custom R scripts.

Each of the population genetic metrics were calculated for the three recent LGT fragments containing multiple genes as well as the 723 fragments of similar size spread across the genome (see above). The latter provided genome-wide distributions, with which LGT values were compared.

The same population genetic statistics were subsequently calculated for each pair of LGT length variants, both within each of the three populations and among pairs of populations for fragments with unambiguous structural variants within populations (LGT_B, LGT_C, and LGT_N).

## DATA AND SOFTWARE AVAILABILITY

The accession for the sequence data reported in this paper is NCBI: PRJNA560360.