



This is a repository copy of *Genome biogeography reveals the intraspecific spread of adaptive mutations for a complex trait*.

White Rose Research Online URL for this paper:
<http://eprints.whiterose.ac.uk/108504/>

Version: Accepted Version

Article:

Olofsson, J.K., Bianconi, M., Besnard, G. et al. (9 more authors) (2016) Genome biogeography reveals the intraspecific spread of adaptive mutations for a complex trait. *Molecular Ecology*, 25 (24). pp. 6107-6123. ISSN 0962-1083

<https://doi.org/10.1111/mec.13914>

This is the peer reviewed version of the following article: Olofsson, J. K. et al (2016), Genome biogeography reveals the intraspecific spread of adaptive mutations for a complex trait. *Mol Ecol.*, which has been published in final form at [Link to final article using the <https://doi.org/10.1111/mec.13914>. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

Reuse

Unless indicated otherwise, fulltext items are protected by copyright with all rights reserved. The copyright exception in section 29 of the Copyright, Designs and Patents Act 1988 allows the making of a single copy solely for the purpose of non-commercial research or private study within the limits of fair dealing. The publisher or other rights-holder may allow further reproduction and re-use of this version - refer to the White Rose Research Online record for this item. Where records identify the publisher as the copyright holder, users can verify any specific terms of use on the publisher's website.

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.



eprints@whiterose.ac.uk
<https://eprints.whiterose.ac.uk/>

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

Journal:	<i>Molecular Ecology</i>
Manuscript ID	MEC-16-1002.R1
Manuscript Type:	Original Article
Date Submitted by the Author:	27-Oct-2016
Complete List of Authors:	Olofsson, Jill; University of Sheffield, Animal and Plant Sciences Bianconi, Matheus; University of Sheffield, Animal and Plant Sciences Besnard, Guillaume; UPS, CNRS, EDB, UMR5174 Dunning, Luke; University of Sheffield, Animal and Plant Sciences Lundgren, Marjorie; University of Sheffield, Animal and Plant Sciences Holota, Helene; Evolution et diversite biologique Vorontsova, Maria; Royal Botanic Gardens Kew Hidalgo, Oriane; Royal Botanic Gardens Kew Leitch, I.J.; Royal Botanic Gardens Kew Nosil, Patrik; University of Sheffield, Animal and Plant Sciences Osborne, Colin; University of Sheffield, Animal and Plant Sciences Christin, Pascal-Antoine; University of Sheffield, Animal and Plant Sciences
Keywords:	Adaptation, C4 photosynthesis, gene flow, lateral gene transfer, Phylogeography

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

Running title: Genome biogeography tracks adaptive mutations

5

Jill K. Olofsson^{1,2}, Matheus Bianconi^{1,2}, Guillaume Besnard³, Luke T. Dunning², Marjorie R. Lundgren², Helene Holota³, Maria S. Vorontsova⁴, Oriane Hidalgo⁴, Ilia J. Leitch⁴, Patrik Nosil², Colin P. Osborne², Pascal-Antoine Christin^{2,5}

10 ¹ These authors contributed equally to the work.

² Department of Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom

³ Laboratoire Évolution & Diversité Biologique (EDB UMR5174), Université de Toulouse, CNRS, ENSFEA, UPS, 118 route de Narbonne, F-31062 Toulouse, France

15 ⁴ Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AB, UK

⁵ Author for correspondence; email: p.christin@sheffield.ac.uk, phone: +44-114-222-0027

Abstract

Physiological novelties are often studied at macro-evolutionary scales such that their micro-
20 evolutionary origins remain poorly understood. Here, we test the hypothesis that key components of
a complex trait can evolve in isolation and later be combined by gene flow. We use C₄
photosynthesis as a study system, a derived physiology that increases plant productivity in warm,
dry conditions. The grass *Alloteropsis semialata* includes C₄ and non-C₄ genotypes, with some
populations using laterally-acquired C₄-adaptive loci, providing an outstanding system to track the
25 spread of novel adaptive mutations. Using genome data from C₄ and non-C₄ *A. semialata*
individuals spanning the species' range, we infer and date past migrations of different parts of the
genome. Our results show that photosynthetic types initially diverged in isolated populations, where
key C₄ components were acquired. However, rare but recurrent subsequent gene flow allowed the
spread of adaptive loci across genetic pools. Indeed, laterally-acquired genes for key C₄ functions
30 were rapidly passed between populations with otherwise distinct gene pools. Thus, our intraspecific
study of C₄-related genomic variation indicates that components of adaptive traits can evolve
separately and later be combined through secondary gene flow, leading to the assembly and
optimization of evolutionary innovations.

35 **Key-words:** adaptation, C₄ photosynthesis, gene flow, lateral gene transfer, phylogeography

Introduction

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

C₄ photosynthesis is a physiological state, present in ~3% of plant species (Sage 2016), which results from the co-ordinated action of multiple enzymes and anatomical components (Hatch 1987; Christin & Osborne 2014). C₄ biochemistry relies on well characterized enzymes that also exist in non-C₄ plants, but with altered abundance, cellular and subcellular localization, regulation, and kinetics (Kanai & Edwards 1999). The main effect of C₄ photosynthesis is an increase in CO₂ concentration at the place of its fixation by the enzyme Rubisco in the Calvin-Benson cycle (von Caemmerer & Furbank 2003). This is advantageous in conditions that restrict CO₂ availability, especially in warm and arid environments under the low-CO₂ atmosphere that has prevailed for the last 30 million years (Sage *et al.* 2012). C₄ plants consequently dominate most open biomes in tropical and subtropical regions, where they achieve high growth rates and large biomass (Griffith *et al.* 2015; Atkinson *et al.* 2016). Despite its apparent complexity, C₄ photosynthesis evolved more than 60 times independently over the ancestral C₃ type (Sage *et al.* 2011), and evolutionary transitions were facilitated by the existence of anatomical and genetic enablers in some groups of plants (Christin *et al.* 2013b; 2015). However, the micro-evolutionary history of photosynthetic transitions is yet to be addressed.

Most C₄ lineages evolved this photosynthetic system millions of years ago, so that the initial changes linked to C₄ evolution remain obscured (Christin & Osborne 2014). In a couple of groups, closely-related species present a spectrum of more or less complete C₄ traits, which is interpreted as the footprint of the gradual C₄ evolution (e.g. McKown *et al.* 2005; Christin *et al.* 2011; Fisher *et al.* 75 2015). These groups provide powerful systems to reconstruct the order of changes during the transition to C₄ photosynthesis (e.g. McKown & Dengler 2007; Heckmann *et al.* 2013; Williams *et al.* 2013). However, the presumed lack of gene flow among these related species impedes testing hypotheses about the importance of secondary gene flow mixing mutations that were fixed in 80 isolated populations. So far, the presence of genotypes with different photosynthetic types has been reported in only one taxon, the grass *Alloteropsis semialata*.

A. semialata includes C₃ and C₄ individuals (Ellis 1974), and a recent study further described individuals with only some of the C₄ anatomical and biochemical components, which allow a weak C₄ cycle (i.e. C₃-C₄ intermediates; Lundgren *et al.* 2016). Other species in this genus, *A. angusta*, *A.* 85 *cimicina*, *A. paniculata*, and *A. papillosa*, are C₄, but perform the C₄ cycle using different enzymes and leaf tissues than *A. semialata*, which points to independent realizations of the C₄ phenotype (Christin *et al.* 2010). Analyses of genes for key C₄ enzymes in a handful of accessions has revealed that some populations of *A. semialata* carry C₄ genes that have been laterally acquired from distant C₄ relatives (Christin *et al.* 2012). The laterally-acquired genes include one for 90 phosphoenolpyruvate carboxykinase (*pck*) and three different copies for phosphoenolpyruvate carboxylase (*ppc*). These laterally-acquired genes are integrated into the C₄ cycle of some extant accessions of *Alloteropsis* (Christin *et al.* 2012; 2013a), but genes for other C₄ enzymes have been transmitted following the species tree (vertically-inherited), and gained their C₄ function via novel mutations (Christin *et al.* 2013a). Some C₄ *Alloteropsis* populations presumably still use the 95 vertically-inherited *ppc* and *pck* homologs for their C₄ cycles. However, the laterally-acquired *ppc* and *pck* copies spent millions of years in other C₄ species, where they acquired adaptive mutations that likely increased their fit for the C₄ function before their transfer (Christin *et al.* 2012). The potential adaptive value of the laterally-acquired genes, as well as their restriction to some C₄ populations, provides a tractable system to elucidate gene movements that led to the emergence and 100 strengthening of the complex C₄ adaptive trait. However, the geographical distributions and frequencies of these laterally-acquired genes are still poorly understood, and the genome history of *A. semialata* remains largely unexplored.

In this study, we obtain low-coverage whole genome sequencing data from *A. semialata* individuals spread across the species' geographical range and differing in photosynthetic type. We 105 use the data to first infer the history of isolation and secondary contact, and then to track the acquisition and spread of the laterally-acquired genes. This biogeographic framework allows us to

test whether the C₄ complex trait was assembled via the sequential fixation of novel mutations within each isolated gene pool or via gene flow combining mutations that had been fixed in distinct gene pools (Fig. 1). In the first scenario, the history of C₄-adaptive mutations, represented by the laterally-acquired genes, would correspond to the sequence of migration and isolation of populations and largely match the history of the rest of the genome (Fig. 1A). In the second scenario, the history of C₄-adaptive mutations would differ from that of the rest of the genome, their selection-driven spread across genetic lineages resulting in more recent coalescence times and gene topologies that differ from the species topology (Fig. 1B). This first intraspecific spatial genomic analysis of key components of the C₄ complex trait opens new avenues to understand the micro-evolutionary processes that led to macro-evolutionary innovations.

Material and Methods

Sampling, sequencing, and genome sizing

A low-coverage whole genome sequencing approach (genome skimming) was used to reconstruct the genome history of *Alloteropsis*. This approach has become increasingly attractive for inferring population parameters (e.g. Buerkle & Gompert 2013; Fumagalli *et al.* 2013) and for studying complex traits (Li *et al.* 2011). It also allows *de novo* assembly of high copy number regions of the genome, such as organelle genomes (Besnard *et al.* 2014; Dodsworth 2015), and can be applied to samples of limited quality and quantity, such as herbarium or museum collections (Besnard *et al.* 2014). Genome skimming data for eleven *A. semialata* individuals, and one of each of the congeneric *A. cimicina* and *A. angusta*, were retrieved from a previous study that used them to assemble chloroplast genomes (Table S1; Lundgren *et al.* 2015, 2016). The photosynthetic type of these samples has been determined previously, and they encompass non-C₄ individuals with and without a weak C₄ cycle, as well as multiple C₄ accessions (Table S1; Lundgren *et al.* 2015, 2016). An additional eight *Alloteropsis* accessions were sampled here to increase the resolution of genome biogeography for the group (Table S1). These include one accession from each of the congeneric species *A. paniculata* and *A. angusta*, and six additional *A. semialata* individuals. These samples were selected to increase the plastid and photosynthetic diversity, with a special focus on the Zambezian biogeographic region (spanning Tanzania, Zambia and the Democratic Republic of Congo – DRC; Linder *et al.* 2012; Table S1), where the majority of the chloroplast and photosynthetic diversities are found (Lundgren *et al.* 2015, 2016). Three of the newly sequenced *A. semialata* accessions ('DRC3', 'TAN3', and 'KEN1') were previously characterized with stable carbon isotopes (Lundgren *et al.* 2015), which can distinguish plants grown using mainly C₄ photosynthesis from those that acquired a significant portion of their carbon via the ancestral C₃ cycle, whether or not it is complemented by a weak C₄ cycle (Smith & Brown 1973; Cerling *et al.*

1997). One of these three accessions ('TAN3') is isotopically intermediate, indicating that a strong C₄ cycle occurs, but that some atmospheric carbon is still fixed directly by the C₃ cycle (Peisker 1986; Monson *et al.* 1988). For four of the new samples, carbon isotopes were measured on a leaf
145 fragment as previously described (Lundgren *et al.* 2015), which revealed that all of them had carbon isotope values within the C₄ range (Table S1).

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

155 The genome size was estimated by flow cytometry for accessions for which live material was available following the one-step protocol of Doležel *et al.* (2007) with minor modifications as described in Clark *et al.* (2016). We selected *Oryza sativa* IR36 (2C=1 pg; Bennett & Smith 1991) and the Ebihara buffer (Ebihara *et al.* 2005) as the most appropriate internal standard and nuclei isolation buffer for all but one accessions (Table S1). For the 'RSA3' accession, whose C-value was
160 estimated to be about three time larger than other accessions, we used the *Pisum sativum* 'Ctirad' standard (2C=9.09 pg; Doležel *et al.* 1992) and the GPB buffer (Loureiro *et al.* 2007), supplemented with 3% of PVP.

Assembly and analyses of chloroplast genomes

165 Complete chloroplast genomes were *de novo* assembled for the newly sequenced individuals using the genome walking method described in Lundgren *et al.* (2015). The newly generated chloroplast genomes were manually aligned with those already available, and a time-calibrated phylogenetic tree was inferred with BEAST v. 1.5.4 (Drummond & Rambaut 2007), as described in Lundgren *et al.* (2015). Monophyly of the outgroup (*A. cimicina* + *A. paniculata*) and the ingroup (*A. angusta* +
170 *A. semialata*) was enforced to root the phylogeny, which is consistent with all previous analyses (Ibrahim *et al.* 2009; Christin *et al.* 2012; GPWGII 2012; Lundgren *et al.* 2015). The root of the tree was fixed to 11 Mya (as found by Lundgren, *et al.* 2015), which was achieved with a normal distribution of mean of 11 and standard deviation of 0.0001. Two different analyses were run for 20,000,000 generations, sampling a tree every 1,000 generations. After checking the convergence of
175 the runs in Tracer v. 1.5.0 (Drummond & Rambaut 2007), the burn-in period was set to 2,000,000

generations, and the maximum credibility tree was identified from the trees sampled after the burn-in period in both analyses, mapping median ages on nodes.

Genotyping across the nuclear genome

180 A reference genome for *Alloteropsis* is currently lacking. However, the grass *Setaria italica* (comment name: Foxtail millet) belongs to the same tribe as *Alloteropsis* (Paniceae) and has a well-assembled reference genome (JGIv2.0.27; Bennetzen *et al.* 2012). *Setaria* and *Alloteropsis* diverged approximately 20 Mya (Christin *et al.* 2012), a time that is sufficient for a complete turn-over of non-coding sequences (Ammiraju *et al.* 2008). However, reads corresponding to coding regions
185 across the genome can still be reliably mapped (see Results).

Raw sequencing reads were quality filtered using the NGS QC toolkit v. 2.3.3 (Patel & Jain 2012). Reads with more than 20% of the bases having a quality score below Q20 and reads with ambiguous bases were removed. Furthermore, low quality bases (<Q20) were trimmed from the 3' end of the remaining reads. The filtered reads were mapped to the *Setaria* reference genome, using
190 Bowtie2 v. 2.2.3 (Langmead *et al.* 2009). Raw alignment files were cleaned using SAMtools v.1.2 (Li *et al.* 2009) and Picard tools v.1.92 (<http://picard.sourceforge.net/>). PCR duplicates were removed, and only uniquely aligned reads in proper pairs were kept. This will remove most of the reads mapped to repetitive sequences, such as transposable elements, while retaining reads mapping to sequences that have been duplicated after the split of *Alloteropsis* and *Setaria*. The cleaned
195 alignments were used to call single nucleotide polymorphic variants (SNPs) with SAMtools v. 0.1.19 using the mpilup function followed by the vcfutil.pl script with default setting supplied with the program. The South African C₄ individual 'RSA3' was excluded during SNP-calling to avoid any bias that might result from the presence of more than two alleles in this polyploid (see Lundgren *et al.* 2015 and Table S1). Genotypes of each accession, including 'RSA3', at all called
200 SNP positions were extracted from the alignments using the mpileup function in SAMtools v.0.1.19, supplying the program with the positions of the called SNPs, and in house developed scripts for further processing (Supplementary Information 1). The low coverage data caused genotype probabilities to be low, which precluded effective filtering based on these probabilities. Therefore, fixed genotype calls were used. To evaluate the proportion of SNPs corresponding to
205 exon sequences, annotations were extracted for the 25,727 coding regions of the *Setaria* genome with homologs in maize and rice genomes (from now on referred to as SZR homologs). The positions of the raw SNPs were intersected with the SZR homolog annotations in BEDTools v.2.19.1 using default settings (Quinlan & Hall 2010).

SNPs with coverage above 2.5 times the genome-wide coverage (Table S2) were converted to
210 unknown genotype calls. Furthermore, genotypes with more than two allele calls were also

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

To test for a bias due to uneven coverage across samples (Table S2), we repeated the phylogenetic analysis on a resampled alignment, where all samples have the same number of bases mapped to the *Setaria* genome. Reads were randomly sampled without replacement from the filtered alignment files until the number of bases across the sampled reads equalled that of the sample with the lowest coverage (Supplementary Information 2). These reanalyses were first conducted with all samples, which resulted in a low number of positions constrained to the samples with the lowest coverage. While analyses on the resampled dataset were consistent with the whole-dataset analyses, the limited number of characters resulted in reduced support. We consequently repeated the resampling allowing for the full alignment of the two *A. semialata* samples with the lowest coverage and alignment success ('AUS1' and 'RSA2') to be retained at a slightly lower coverage than the rest of the samples. SNPs were called as outlined above, which allowed for the retention of 22,821 SNPs.

230

Genetic structure and test for secondary gene flow

Preliminary cluster analyses with a focus on *A. semialata* showed that a more stringent filtering of the SNPs improved convergence of the analyses. Only positions with less than 10% missing data (2,607 SNPs) within this species were consequently kept for analyses of its population structure, using the STRUCTURE software (Pritchard *et al.* 2000). Ten independent analyses were run for each number of population components (K) from one to ten, under the admixture model. The adequate run length and burn-in periods were determined through preliminary analyses, which indicated that a burn-in period of 300,000 generations followed by 200,000 iterations provided stable estimates for all K values. The optimal K values were determined using the method of Evanno *et al.* (2005), as implemented in STRUCTUREHarvester (Earl & von Holdt 2012). The results of the ten runs for each K were summarized using CLUMPP v. 1.1.2 (Jakobsson & Rosenberg 2007) and graphically displayed using DISTRUCT v. 1.1 (Rosenberg 2004). These analyses were repeated without the polyploid individual 'RSA3', which led to the same cluster assignments, showing that differences in ploidy levels do not affect the conclusions. Finally, the

240

245 cluster analyses were repeated on alignments based on the reads subsampled to similar coverage in all sampled (see above), allowing for 25% missing data per site (retention of 681 SNPs).

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

Assembly and analyses of selected genes

Two different groups of closely related genes were selected for detailed analyses. The genes selected were two C₄-related protein-coding genes, phosphoenolpyruvate carboxylase (*ppc* genes) and phosphoenolpyruvate carboxykinase (*pck* genes), that include some copies acquired by *Alloteropsis* from distantly related species via lateral gene transfer, while other copies were vertically inherited following the species tree (Christin *et al.* 2012). Previous conclusions regarding the distribution of these genes among accessions of *Alloteropsis* were based on PCR and Sanger sequencing, which can be biased due to the possibility of primer binding mismatches. The presence/absence of the laterally-acquired *ppc* and *pck* genes and their vertically-inherited homologs across the accessions were therefore re-evaluated here using the genome-skimming data, as well as new PCR and Sanger sequencing with primer verified against the new genomic data. Using molecular dating, the divergence times of the laterally-acquired genes were compared to those of vertically-inherited homologs belonging to the same set of accessions.

275 Reads were first mapped on gene segments of the *ppc* and *pck* genes from different accessions of *Alloteropsis* (grass co-orthogols *ppc-1P3* and *pck-1P1*; Christin *et al.* 2012, 2015). These segments have been previously sequenced and analysed in a number of other C₃ and C₄ grasses (Christin *et al.* 2012). The availability of this rich reference dataset allows mapping to closely related accessions of *Alloteropsis*, which improves the alignment success compared to the whole

280 genome approach described above, and increases the confidence in the assignment. The gene segments cover exons 8 to 10 for *ppc* and exons 3 to 10 for *pck*, including introns, and represent approximately 46% (1,492 bp) and 63% (1,487 bp), respectively, of the full length coding sequences. In-house Perl scripts (Supplementary Information 3) were used to unambiguously assign reads to genes of these datasets, following the phylogenetic annotation method of Christin *et al.* (2015). In summary, this approach consists of: (i) building a reference dataset of sequences with known identity for closely-related gene lineages, (ii) using blast searches to identify all sequences homologous to any of these reference sequences in the query dataset (the filtered reads in this case), (iii) independently aligning each homologous sequence to the reference dataset and inferring a phylogenetic tree, and (iv) establishing the identity of each of the query sequences based on the phylogenetic group in which it is nested. Assignment of reads to the gene lineages was verified by visual inspection of the phylogenetic trees and the alignments. Subsequently, all reads assigned to each of the vertically-inherited or laterally-acquired gene lineages were retrieved, and aligned to PCR-isolated sequences (see Results) using Geneious v. 6.8 (Kearse *et al.* 2012). The reads were then assembled into gene models, comprising introns and exons, for the studied segments. Multiple gene models were assembled for a single individual when the existence of distinct alleles was supported by at least two different polymorphic sites, each with at least two independent reads. Paired-end reads were then merged into contigs if they shared the polymorphisms. Reads that did not overlap the polymorphic sites were merged with all alleles, replacing additional polymorphisms with IUPAC ambiguity codes.

300 To check whether partial pseudogenes that do not include the studied segments exist in some genomes, the presence of laterally-acquired *ppc* genes was also tested using only coding sequences corresponding to exons 1 to 7, which were retrieved from a transcriptome study of *Alloteropsis semialata* (Christin *et al.* 2012). This transcriptome was generated for a South African C₄ polyploid with two laterally-acquired *ppc* genes, but the vertically-inherited versions of *ppc* and *pck* were not available in this transcriptome, preventing phylogenetic analyses. Blastn searches were used to identify reads mapping to one of the two laterally-acquired *ppc* genes on at least 50 bp with at least 99% of identity. Finally, the presence/absence of the different *pck* and *ppc* copies was further confirmed via PCR and Sanger sequencing using primers specific to the different gene copies (Table S3; Christin *et al.* 2012). PCR, purification and sequencing were conducted as described in Lundgren *et al.* (2015), except for changes of the annealing temperature and/or extension time (Table S3). These PCR were conducted only on samples for which good quality DNA was available. Indeed, DNA isolated from herbarium samples is highly degraded, precluding reliable PCR screening.

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

325

Results

Read alignment and SNP calling

The number of filtered paired-end reads varied across samples, for a genome-wide coverage ranging
from 0.70 to 4.52 (Table S2). The proportion of filtered paired-end reads that aligned to the *Setaria*
330 genome varied between 4.04 and 10.23%, and between 1.22 and 2.94% aligned to the coding
regions (Table S2). While the mapping was done across the whole genome of *Setaria* (excluding the
organelle genomes), high divergence of non-coding sequences means that high mapping success is
expected to be concentrated mostly onto coding sequences. About 9% of the *Setaria* genome
corresponds to exons (Bennetzen *et al.* 2012). Assuming that the total length of exons is similar in
335 the two species, the larger genome of *Alloteropsis* means that this proportion should be about 4.5%,
so that roughly half of the reads corresponding to nuclear exons were mapped. The rest of the reads
that belong to exons probably correspond to gene sections that are too divergent between the two
species to successfully map.

Only uniquely aligned reads were used to call SNPs, which inherently excludes common
340 repetitive regions such as transposons. However, 1,111 raw SNPs had a higher than expected
coverage (>5x) across at least 50% of the samples. The positions of 91% (1,007) of these high
coverage SNPs fell outside of the SZR homolog regions and the rest were concentrated to only 14
SZR homologs. We therefore hypothesize that these high coverage SNPs stem from genetic regions
(mostly non-coding) that have been duplicated after the split between *Alloteropsis* and *Setaria* and
345 they were subsequently removed from the analyses.

A total of 170,629 SNPs with less than 50% missing data across the 21 accessions were
finally selected for downstream analyses. These sites are spread across all chromosomes (Fig. S1)
and 96% of them fall within one of 9,948 SZR homologs. The 2,607 SNPs used for the cluster

analysis (less than 90% missing data across the *A. semialata* samples) were equally well spread
350 across the genome (Fig. S1) and 97% fall within one of 848 SZR homologs. Most of the variation in
missing data across samples (Table S2) is likely explained by differences in coverage, although
presence/absence of genes within each accession might also influence the individual mapping
success.

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

Phylogenetic trees

The plastid phylogeny identified two C₄ individuals from DRC with a haplotypes that form a new
360 C₄ plastid lineage based on divergence times (i.e. lineage G, sister to lineage F; Figs 2 and S2).
Relationships based on markers sampled across the nuclear genome confirm the monophyly of *A.*
semialata and its sister-group relationship to *A. angusta*, but present multiple incongruences with
the chloroplast tree within *A. semialata* (Figs 2 and S3). In this genome-wide tree, the first
divergence leads to a group composed of the non-C₄ accessions of *A. semialata* from South Africa
365 without any known C₄ cycle (Clade I; Fig 2 and S3), and the second divergence leads to a group
comprising the non-C₄ accessions from the Zambezian region that use a weak C₄ cycle (Clade II;
Fig 2 and S3; C₃-C₄ intermediates *sensu* Lundgren *et al.* 2016). The isotopically intermediate
accession 'TAN3' is then sister to all C₄ accessions (Figs 2 and S3). The two C₄ accessions bearing
the plastid lineage G form a paraphyletic clade, while the other C₄ accessions from the Zambezian
370 region ('TAN4', 'DRC3', and 'DRC4') are grouped in a strongly supported clade (Clade III; Figs 2
and S3). The South African polyploid individual 'RSA3' is sister to the C₄ individuals sampled
outside of the Zambezian region and the rest of the C₄ accessions form the strongly supported clade
IV, with two subclades corresponding to Africa plus Madagascar and Asia plus Australia (Figs 2
and S3). The nuclear phylogeny based on the resampled dataset is mostly identical to the one based
375 on the whole dataset (Figs S3 and S4).

Genetic structure and secondary gene flow within A. semialata

Based on the whole genome clustering analysis, four clusters explain the dataset best, and adding
groups does not improve the likelihood (Fig. 3B). However, the method of Evanno *et al.* (2005)
380 indicates that the maximum fit improvement is at two clusters, with four clusters representing the
second maximum fit improvement (Fig. 3C). With four clusters, the main clades from the whole
genome phylogeny are recovered (Figs 2 and 3A). This genetic structure matches the photosynthetic
types rather than the geographic origin, with the non-C₄ clades I and II and the C₄ clades III and IV

each forming distinct homogenous groups (Fig. 3A). The three Zambebian individuals that formed a
385 paraphyletic clade in the nuclear phylogeny ('TAN3', 'DRC1', and 'DRC2') are partially assigned to
two Zambebian groups, the non-C₄ clade II and the C₄ clade III (Figs 2 and 3A). Finally, the
polyploid individual from South Africa, 'RSA3', is partially assigned to the two C₄ clades III and IV
(Fig. 3A). The cluster results based on the resampled dataset are less stable due to a lower number
of sites, but the assignments are similar (Figs 3 and S5).

390 Heterozygosity was estimated for each sample based on the 22,821 SNPs from the resampled
dataset with similar coverage across samples. While these estimates are based only on sites variable
within *Alloteropsis* and should consequently not be interpreted as genome-wide heterozygosity, it is
possible to compare the estimates between samples. The individuals assigned to multiple clusters
had the highest percentage of heterozygous SNPs (Fig. S6), which confirms their genetic diversity.

395 Together, our intraspecific genetic analyses reveal the existence of distinct gene pools despite
overlapping distributions (Figs 3A and 4), but also suggest that genetic exchanges have happened
among groups. The incongruences between the phylogenetic structures of the chloroplast and
nuclear genomes, together with the assignment of some individuals to multiple genetic clusters,
suggest that the three Zambebian individuals 'TAN3', 'DRC2', and 'DRC1' have ancestors from
400 distinct genetic groups, in this case the nuclear clades II and III. ABBA-BABA tests were therefore
conducted to test this hypothesis, using *A. angusta* (individual Ang2) as the outgroup. The
individual 'TAN4' was selected as the representative of clade III because it is geographically more
distant and distinct on all genetic markers (Figs 4, S2, and S3). Significant indications ($P < 0.05$ after
correction for multiple testing) of gene flow from the non-C₄ clade II ('TAN2' and 'TAN1') into the
405 populations assigned to multiple clusters ('TAN3', 'DRC2', and 'DRC1') were found (Table 1). In
contrast, there is no evidence of a significant secondary contribution of clade II into individuals of
clade III ('DRC3' or 'DRC4'; Table 1). However, in one case, a slight excess of BABA sites was
detected, which was not significant after correction for multiple testing (Table 1). This would
suggest some genetic contribution from one non-C₄ population of clade II ('TAN1') into the C₄
410 population represented by 'TAN4' (Table 1).

Within clade IV, the C₄ individual from Madagascar ('MAD1') was grouped with Asian C₄
accessions on plastid genomes but grouped with the African C₄ accessions based on markers from
across the nuclear genome (Figs S2 and S3). An ABBA-BABA test was conducted to test the
hypothesis of secondary gene flow after the split of the African and Asian C₄ accessions. The
415 accession 'TAN4' was used as the outgroup, being sister to all accessions from clade IV. The
Taiwan accession ('TPE1') was selected as the Asian sample, while the Burkinabe accession
('BUR1') represented Africa. Overall more ABBA than BABA sites were detected (Table 1),
indicating that the Asian accession was closer to the accession from Madagascar ('MAD1') than to

the accession from mainland Africa, but the D statistic for this test was not significant after
420 correcting for multiple testing (Table 1). Plastid markers, which represent seed dispersal, group the
Madagascan accessions with Asian individuals. Therefore, a possible scenario involves an initial
seed dispersal from Africa to Madagascar and then from Madagascar to Asia, with subsequent
pollen flow between Africa and Madagascar.

425 *Assembly and analyses of selected genes*

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

Together, the results confirmed previous phylogenetic analyses (Christin *et al.* 2012), but with
430 a significant increase of the sample size. Reads assigned to the *pck* gene copy laterally acquired
from members of the *Cenchrus* genus (*pck-IP1_LGT:C*) were detected in the two *A. angusta*
accessions and all *A. semialata* accessions, except the two non-C₄ accessions from South Africa
(Table 2; Figs S7 and S8). The sequences assembled for the laterally-acquired *pck* gene were highly
similar between the different accessions, leading to a poorly resolved phylogeny (Fig. S8). By
435 contrast, the sequences assembled for the vertically-inherited *pck* gene were variable among
accessions, and the nuclear clades I and II were recovered in their phylogeny, while clades III and
IV were not differentiated (Fig. S8). Interestingly, one accession with mixed genetic backgrounds
(‘DRC1’) has two divergent alleles, one of which groups with clade II and the other with clade
III/IV (Fig. S8). Dating analyses indicate that the divergence of *A. angusta* and *A. semialata* is more
440 recent for the laterally-acquired *pck* than for the vertically-inherited copy (Fig. S9). However, the
divergence of C₄ accessions of *A. semialata* is estimated at a similar time based on the vertically-
inherited and laterally-acquired *pck* (Fig. 5).

The vertically-inherited *ppc* was recovered from all samples, and the assembled gene models
were variable enough to partially resolve the phylogeny, with well supported clades corresponding
445 to the different species (Fig. S10). Although support was limited within *A. semialata*, the non-C₄
clades I and II (including sequences from individuals assigned to multiple clades) were sister to a
clade composed of the C₄ accessions from clade IV nested within those of clade III (Fig. S10). The
divergence of vertically-inherited *ppc* from C₄ accessions (excluding those partially assigned to
clusters other than III and IV) matches the divergence of the vertically-inherited *pck* for the same
450 accessions (Fig. 5).

The *ppc* gene laterally acquired from Andropogoneae (*ppc-IP3_LGT:A*) was only detected in
the Australian C₄ accession (‘AUS1’; Table 2, Figs S7 and S10). An almost complete sequence for
the studied segment was assembled, which was identical to those isolated by PCR.

The *ppc* gene laterally acquired from the *Setaria palmifolia* complex (*ppc-IP3_LGT:C*) was
455 detected in the C₄ accession from South Africa ('RSA3') and the DRC (Table 2, Figs S7 and S10).
Although no reads matching exons 8-10 of *ppc-IP3_LGT:C* was detected in the accession 'TAN3', a
total of seven reads from this individual matched exons 1-7. This suggests that the gene is truncated
and probably exists as a pseudogene in this individual. The *ppc-IP3_LGT:C* sequences were largely
460 conserved, although distinct alleles were assembled in one of the accessions with mixed genetic
background ('DRC2'; Fig. S10). The divergence of *ppc-IP3_LGT:C* genes belonging to different C₄
accessions was more recent than for the vertically-inherited *ppc* and *pck* of the same accessions
(Fig. 5).

The *ppc* gene acquired from Melinidinae (*ppc-IP3_LGT:M*) was identified in nine C₄
accessions of *A. semialata*, the isotopically intermediate *A. semialata*, and the two congeners *A.*
465 *cimicina* and *A. paniculata* (Table 2, Figs S7 and S10). Highly divergent alleles of the *ppc-*
IP3_LGT:M gene were inferred for *A. cimicina* and *A. paniculata* (Fig. S10). However, the
sequences of *ppc-IP3_LGT:M* from *A. semialata* were very similar to each other, and nested within
the alleles from *A. cimicina/paniculata* (Fig. S10). The split of *A. semialata* and *A. cimicina* is more
recent for *ppc-IP3_LGT:M* than for the vertically-inherited *ppc* and *pck* (Fig. S9). In addition, the
470 divergence of C₄ accessions of *A. semialata* based on this *ppc-IP3_LGT:M* gene occurred more
recently than the divergence based on the vertically-inherited *ppc* and *pck* (Fig. 5).

Discussion

Divergence of photosynthetic types in isolation followed by secondary gene flow

475 Overall, our genome-wide analyses reveal a strong genetic structure, which matches photosynthetic
types between than geographic origins, although both play a role. All C₄ individuals form a
monophyletic group based on genome-wide markers, which is sister to a clade composed of non-C₄
accessions from the Zambebian region with a weak C₄ cycle (clade II; Figs 2 and 4), and together
these two groups are sister to the non-C₄ accessions from South Africa that lack a C₄ cycle (clade I;
480 Figs 2 and 4). The C₄ clade contains two clearly distinct subgroups, one from the Zambebian region
(clade III; Figs 2 and 4) and the other one encompassing all C₄ accessions sampled outside this
region, from Western Africa to Australia (clade IV; Figs 2 and 4). The Zambebian region
encompasses more genetic diversity than the rest of the species' range, including a total of five
plastid lineages, four of which are endemic (clades B, C, F, and G; Figs 2 and S2). This finding
485 further supports this region as the centre of origin for *A. semialata* (Lundgren *et al.* 2015). Based on
both plastid and nuclear genomes, the divergence of photosynthetic types likely also happened
within this region (Fig. 6). Both C₄ and non-C₄ populations in the Zambebian region are associated
with Miombo woodlands. Periodic cycles of contraction and expansion of these wooded savannas

during recent geological times might have isolated populations of *A. semialata* in this geologically
490 and topographically complex region (Cohen *et al.* 2007; Beuning *et al.* 2011). The ancestral
photosynthetic state is likely non-C₄ and mutations altering the leaf anatomy and upregulation of
enzymes already present in the non-C₄ ancestors likely led to the emergence of a constitutive C₄
cycle in some isolated populations (Mallmann *et al.* 2014; Bräutigam & Gowik 2016). One of the
lineages descending from the initial C₄ pool, corresponding to clade IV, later left the Miombo of the
495 Zambezi region and rapidly spread across Africa and all the way to Asia and Australia (Figs 4
and 6). This biogeographical history therefore points to the initial emergence of the C₄ physiology
in *A. semialata* within the Zambezi region, with subsequent isolation of the C₄ descendants (Fig.
6).

The lack of association between chloroplast and nuclear groups (Figs 2, S2, and S3) in the
500 Zambezi region suggests ancient, but recurrent, secondary gene flow followed by
homogenization of the local gene pools. In addition, the presence of three individuals with mixed
nuclear backgrounds indicates relatively recent gene flow between previously isolated groups. The
maximum expansion of the Miombo woodlands during interglacial periods, as presently occurs,
would likely favour seed dispersal over a larger area, leading to secondary contacts (Vincens 1989;
505 Cohen *et al.* 2007; Beuning *et al.* 2011), a process frequently reported in Europe (reviewed in e.g.
Hewitt 2000; Schmitt 2007). We propose that this expansion allowed genetic exchanges between
previously isolated lineages, some of which had made the transition to a full C₄ physiology during
the previous isolation. No evidence of gene flow between C₄ and non-C₄ individuals was found
outside of the Zambezi region, and crosses might be prevented in South Africa, the other region
510 where C₄ and non-C₄ individuals overlap, by differences in ploidy levels (Fig. 4; Liebenberg &
Fossey 2001). However, our analyses suggest that allopolyploidy contributed to the mixing of
nuclear groups III and IV in Southern Africa (Fig. 6). In addition, while the recent divergence
decreases statistical confidence, we found suggestions for secondary gene flow between different
subgroups of the C₄ clade IV in Madagascar (Fig. 6).

515 Repeated isolation followed by recurrent, but rare secondary gene flow has created a dynamic
population structure whereby adaptive mutations, such as those for the C₄ trait, can appear and
sweep to fixation in isolation and later come together through admixing in times of contact. While
mutations for increasing the expression of the C₄-related genes and altering the leaf anatomy are
unknown, genes for two key C₄ enzymes were laterally acquired by *A. semialata* (Christin *et al.*
520 2012). These lateral gene transfers likely took place in *A. semialata* plants that already used C₄
photosynthesis, and once acquired, these genes presumably replaced the function of the vertically-
inherited gene copies that were overexpressed but not biochemically optimized (Christin *et al.*
2012). The biogeographic history inferred here for the nuclear genome allows us to estimate the

region where these lateral gene transfers likely occurred and track the subsequent spread of these
525 genes among different gene pools.

Spread of C₄-adaptive mutations among gene pools

Our analyses detected the laterally-acquired *pck* gene in all *A. angusta* and *A. semialata* individuals
apart from two non-C₄ *A. semialata* South African accessions, confirming previous PCR-based
530 approaches (Table 2; Christin *et al.* 2012). The divergence time is younger between the laterally-
acquired *pck* genes from *A. angusta* and *A. semialata* than between the vertically-inherited genes of
the same species (Figs 5 and S9). This suggests that the laterally-acquired *pck* was passed between
A. angusta and *A. semialata* through secondary gene flow.

The accessions from Taiwan and Thailand do not possess any laterally-acquired *ppc* genes,
535 yet carbon isotopes unambiguously indicate that they carry out C₄ photosynthesis (Table 2;
Lundgren *et al.* 2015). It is therefore likely that they overexpress their vertically-inherited *ppc* and
other genes required to generate a working C₄ cycle in the absence of repeated rounds of fixation of
adaptive amino acids, as observed in older C₄ lineages (Christin *et al.* 2007; Besnard *et al.* 2009;
Huang *et al.* 2017).

540 Out of the three different *ppc* genes acquired via lateral gene transfers from distant C₄
relatives (Table 2; Christin *et al.* 2012), only *ppc-IP3_LGT:A* is restricted to one of the accessions
sampled here ('AUS1'). This gene was only found in Australia, and it is thus likely that it was
recently acquired in this region (Fig. 6). The other two laterally-acquired *ppc* genes are absent from
some individuals, but spread across multiple populations of *A. semialata* that belong to different
545 genomic clusters (Table 2). This pattern could result from the presence of the gene in the common
ancestor and subsequent losses in some populations. However, this scenario is not supported by the
lack of phylogenetic structure on the laterally-acquired genes (Fig. S10) and the comparison of
divergence times, which indicate that the divergence of variants of both *ppc-IP3_LGT:M* and *ppc-
IP3_LGT:C* found in C₄ accessions is more recent than the divergence of vertically-inherited genes
550 in the same accessions (Fig. 5).

The laterally-acquired *ppc-IP3_LGT:M* gene was identified in the C₄ congeners *A. cimicina*
and *A. paniculata*, as well as all C₄ accessions of *A. semialata* from Africa and Madagascar
(whether from clade III or IV; Table 2). However, this gene was absent from the Asian/Australian
C₄ accessions from clade IV and the African non-C₄ (clades I and II; Table 2). The divergence time
555 between *ppc-IP3_LGT:M* genes belonging to *A. cimicina* and *A. semialata* is younger than the
divergence times for the vertically-inherited genes from the same species (Fig. S9). In addition, the
higher allelic diversity in *A. cimicina* compared to *A. semialata* suggests that the *ppc-IP3_LGT:M*
gene was first acquired by *A. cimicina* and then transferred to *A. semialata*, potentially via

hybridization. This gene has subsequently likely spread across distinct genetic groups of *A. semialata* in Africa and Madagascar via secondary pollen flow (Fig. 6). The fixation of the *ppc-IP3_LGT:M* gene within different populations would have been favoured by its improvement of the C₄ cycle, a function for which it was already optimized after millions of years spent in another C₄ lineage. Once this adaptive gene copy was acquired in a population, the vertically-inherited *ppc* copy probably underwent pseudogeneization as a result of relaxed selection. Indeed, the vertically-inherited *ppc* genes bear frameshift mutations causing loss of function in two accessions with the laterally-acquired *ppc-IP3_LGT:M* ('TAN4' and 'Cim1'), supporting the hypothesis that their function was taken over by the newly acquired gene, making them obsolete.

The last of the laterally-acquired *ppc* genes, *ppc-IP3_LGT:C*, was found in the South African C₄ polyploid ('RSA3') as well as in four C₄ and one isotopically intermediate individuals from the Zambezi region, two from clade III and three with genetic contributions from clades II and III (Table 2). This gene was laterally acquired from a species of the *Setaria palmifolia* complex (Christin *et al.* 2012), which co-occurs with *A. semialata* in Zambezi Africa, where they grow metres apart, but not in South Africa (Clayton 1979). The transfer therefore likely occurred in the Zambezi region and later spread among the C₄ populations in this region through secondary gene flow (Fig. 6). Once acquired the *ppc-IP3_LGT:C* gene presumably took over the C₄ function, which might have been fulfilled by the previously acquired *ppc-IP3_LGT:M*. Indeed, *ppc-IP3_LGT:M* is still expressed in the transcriptome of the South African C₄ accession, but possesses internal stop codons that prevent proper translation (Christin *et al.* 2012). The newly acquired *ppc-IP3_LGT:C* likely spread to the C₄ populations from South Africa, through the putative segmental allopolyploidy event, providing a mechanism to propagate adaptive loci across genetic pools (Fig. 6). However, the Melinidinae *ppc-IP3_LGT:M* discussed above was spread among diploid individuals from clades III and IV, showing that adaptive loci can be transmitted despite limited gene flow, without the need for polyploidization.

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

Conclusions

595 In this study, we analysed genomic data from multiple accessions of the grass *A. semialata* using low-coverage whole genome sequencing. Using a biogeographic framework for different parts of the genome, we demonstrate that multiple genetic pools exist, which are generally associated with different photosynthetic types. These pools originated more than 2 million years ago in the Zambebian region and were kept relatively isolated, but with recurrent secondary gene flow, including between non-C₄ and C₄ individuals. These genetic exchanges contributed to the spread of adaptive loci, as illustrated by key C₄ genes acquired laterally in the Zambebian region and then rapidly passed to other African C₄ accessions. This process likely gradually optimized the initial C₄ pathway of some *A. semialata* populations through the assembly of different components. These genetic elements evolved in different parts of the species range, where limited gene flow might have facilitated local adaptation, but their subsequent combination likely improved the efficiency of the photosynthetic pathway of some accessions.

Acknowledgements

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

References

- Ammiraju JSS, Lu F, Sanyal A, *et al.* (2008). Dynamic evolution of *Oryza* genomes is revealed by comparative genomic analysis of a genus-wide vertical data set. *Plant Cell*, **20**, 3191-3209.
- 620 Atkinson RRL, Mockford EJ, Bennett C, *et al.* (2016) C₄ photosynthesis boost growth via altered physiology, allocation and size. *Nature Plants*, **2**, 16038.
- Bennett MD, Smith JB (1991) Nuclear DNA amounts in angiosperms. *Philosophical Transactions of the Royal Society of London. Series B*, **334**, 309-345.
- 625

Bennetzen JL, Schmutz J, Wang H, *et al.* (2012) Reference genome sequence of the model plant *Setaria*. *Nature Biotechnology*, **30**, 555-561.

630 Besnard G, Christin PA, Male PJ, *et al.* (2014) From museums to genomics: Old herbarium specimens shed light on a C₃ to C₄ transition. *Journal of Experimental Botany*, **65**, 6711-6721.

Besnard G, Muasya AM, Russier F, Roalson EH, Salamin N, Christin PA (2009) Phylogenomics of C₄ photosynthesis in sedges (Cyperaceae): multiple appearances and genetic convergence.

635 *Molecular Biology and Evolution*, **26**, 1909-1919.

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

640

Bräutigam A, Gowik U (2016) Photorespiration connects C₃ and C₄ photosynthesis. *Journal of Experimental Botany*, **67**, 2953-2962.

Buerkle CA, Gompert ZA (2013) Population genomics based on low coverage sequencing: how low should we go? *Molecular Ecology*, **22**, 3028-3035.

645

Cerling TE, Haris JM, MacFadden BJ, *et al.* (1997) Global vegetation change through the Miocene/Pliocene boundary. *Nature*, **389**, 153-158.

650 Christin PA, Arakaki M, Osborne CP, Edwards EJ (2015) Genetic enablers underlying the clustered evolutionary origins of C₄ photosynthesis in angiosperms. *Molecular Biology and Evolution*, **32**, 846-858.

Christin PA, Boxall SF, Gregory R, Edwards EJ, Hartwell J, Osborne CP (2013a) Parallel
655 recruitment of multiple genes into C₄ photosynthesis. *Genome Biology and Evolution*, **5**, 2174-2187.

Christin PA, Edwards EJ, Besnard G, *et al.* (2012) Adaptive evolution of C₄ photosynthesis through recurrent lateral gene transfer. *Current Biology*, **22**, 445-449.

660

Christin PA, Freckleton RP, Osborne CP (2010) Can phylogenetics identify C₄ origins and reversals? *Trends in Ecology & Evolution*, **25**, 403-409.

665 Christin PA, Osborne CP (2014) The evolutionary ecology of C₄ plants. *New Phytologist*, **204**, 765-781.

Christin PA, Osborne CP, Chatelet DS, *et al.* (2013b) Anatomical enablers and the evolution of C₄ photosynthesis in grasses. *Proceedings of the National Academy of Sciences*, **110**, 1381-1386.

670 Christin PA, Sage TL, Edwards EJ, Ogburn RM, Khoshravesh R, Sage RF (2011) Complex evolutionary transitions and the significance of C₃-C₄ intermediate forms of photosynthesis in Molluginaceae. *Evolution*, **65**, 643-660.

675 Christin PA, Salamin N, Savolainen V, Duvall MR, Besnard G (2007) C₄ photosynthesis evolved in grasses via parallel adaptive genetic changes. *Current Biology*, **17**, 1241-1247.

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

680 Clayton WD (1979) Notes on *Setaria* (Gramineae). *Kew Bulletin*, **33**, 501-509.

Cohen AS, Stone JR, Beuning KR, *et al.* (2007) Ecological consequences of early late pleistocene megadroughts in tropical Africa. *Proceedings of the National Academy of Science*, **104**, 16422-16427.

685 Dodsworth S (2015) Genome skimming for next-generation biodiversity analysis. *Trends in Plant Science*, **20**, 525-527.

690 Doležel J, Greilhuber J, Suda J (2007) Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols*, **2**, 2233-2244.

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

- 695 Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology*, **7**, 214.
- Durand EY, Patterson N, Reich D, Slatkin M (2011) Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, **28**, 2239-2252.
- 700 Earl DA, vonHoldt BM (2012) STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, **4**, 359-361.
- 705 Ebihara A, Ishikawa H, Matsumoto S, *et al.* (2005) Nuclear DNA, chloroplast DNA, and ploidy analysis clarified biological complexity of the *Vandenboschia radicans* complex (Hymenophyllaceae) in Japan and adjacent areas. *American Journal of Botany*, **92**, 1535-1547.
- Edgar RC (2004) MUSCLE: A multiple sequence alignment method with reduced time and space
710 complexity. *BMC Bioinformatics*, **5**, 113.
- Ellis RP (1974) The significance of the occurrence of both Kranz and non-Kranz leaf anatomy in the grass species *Alloteropsis semialata*. *South African Journal of Science*, **70**, 169-173.
- 715 Ellstrand NC (2014) Is gene flow the most important evolutionary force in plants? *American Journal of Botany*, **101**, 737-753.
- Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Molecular Ecology*, **14**, 2611-2620.
- 720 Fisher AE, McDade LA, Kiel CA, *et al.* (2015) Evolutionary history of *Blepharis* (Acanthaceae) and the origin of C₄ photosynthesis in section *Acanthodium*. *International Journal of Plant Sciences*, **176**, 770-790.
- 725 Fumagalli M, Vieira FG, Korneliussen TS, *et al.* (2013) Quantifying Population Genetic Differentiation from Next-Generation Sequencing Data. *Genetics*, **195**, 979-992.
- Geber MA, Griffen LR (2003) Inheritance and natural selection on functional traits. *International Journal of Plant Sciences*, **164**, S21-S42.

730

GPWG, II (2012) New grass phylogeny resolves deep evolutionary relationships and discovers C₄ origins. *New Phytologist*, **193**, 304-312.

735

Green RE, Krause J, Briggs AW, *et al.* (2010) A draft sequence of the Neandertal genome. *Science*, **328**, 710-722.

740

Griffith DM, Anderson TM, Osborne CP, Strömberg CAE, Forrestel EJ, Still CJ (2015) Biogeographically distinct controls on C₃ and C₄ grass distributions: Merging community and physiological ecology. *Global Ecology and Biogeography*, **24**, 304-313.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, **59**, 307-321.

745

Hatch MD (1987) C₄ photosynthesis: A unique blend of modified biochemistry, anatomy and ultrastructure. *Biochimica et Biophysica Acta*, **895**, 81-106.

Heckmann D, Schulze S, Denton A, *et al.* (2013) Predicting C₄ photosynthesis evolution: Modular, individually adaptive steps on a Mount Fuji fitness landscape. *Cell*, **153**, 1579-1588.

750

Hewitt G (2000) The genetic legacy of the Quaternary ice ages. *Nature*, **405**, 907-913.

Huang P, Studer AJ, Schnable JC, Kellogg EA, Brutnell TP (2017) Cross species selection scans identify components of C₄ photosynthesis in the grasses. *Journal of Experimental Botany*, doi:10.1093/jxb/erw256.

755

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

760

Ibrahim DG, Burke T, Ripley BS, Osborne CP (2009) A molecular phylogeny of the genus *Alloteropsis* (Panicoidae, Poaceae) suggests an evolutionary reversion from C₄ to C₃ photosynthesis. *Annals of Botany*, **103**, 127-136.

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

Kanai R, Edwards GE (1999) The biochemistry of C4 photosynthesis. In: C4 Plant Biology (eds
Sage RF, Monson RK), pp. 49–87. Academic Press.

770

Kearse M, Moir R, Wilson A, *et al.* (2012) Geneious basic: An integrated and extendable desktop
software platform for the organization and analysis of sequence data. *Bioinformatics*, **28**, 1647-
1649.

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

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

780

Leinonen T, Cano JM, Makinen H, Merila J (2006) Contrasting patterns of body shape and neutral
genetic divergence in marine and lake populations of threespining sticklebacks. *Journal of*
Evolutionary Biology, **19**, 1803-1812.

785 Lenski RE, Ofria C, Pennock RT, Adami C (2003) The evolutionary origin of complex features.
Nature, **423**, 139-144.

Li H, Handsaker B, Wysoker A, *et al.* (2009) The sequence alignment/map format and SAMtools.
Bioinformatics, **25**, 2078-2079.

790

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

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

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

800

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

Lundgren MR, Besnard G, Ripley BS, *et al.* (2015) Photosynthetic innovation broadens the niche within a single species. *Ecology Letters*, **18**, 1021-1029.

805

Lundgren MR, Christin PA, Gonzalez Escobar E, *et al.* (2016) Evolutionary implications of C₃-C₄ intermediates in the grass *Alloteropsis semialata*. *Plant Cell & Environment*, **39**, 1874-1885

Mallmann J, Heckmann D, Bräutigam A, *et al.* (2014) The role of photorespiration during the evolution of C₄ photosynthesis in the genus *Flaveria*. *eLife*, **3**, e02478.

810

McKown AD, Dengler NG (2007) Key innovations in the evolution of Kranz anatomy and C₄ vein pattern in *Flaveria* (Asteraceae). *American Journal of Botany*, **94**, 382-399.

815

McKown AD, Moncalvo JM, Dengler NG (2005) Phylogeny of *Flaveria* (Asteraceae) and inference of C₄ photosynthesis evolution. *American Journal of Botany*, **92**, 1911-1928.

Meléndez-Hevia E, Waddell TG, Cascante M (1996) The puzzle of the Krebs citric acid cycle: assembling the pieces of chemically feasible reactions, and opportunism in the design of metabolic pathways during evolution. *Journal of Molecular Evolution*, **43**, 293-303.

820

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

825

Monson RK, Teeri JA, Ku MS, Gurevitch J, Mets LJ, Dudley S (1988) Carbon-isotope discrimination by leaves of *Flaveria* species exhibiting different amounts of C₃- and C₄-cycle co-function. *Planta*, **174**, 145-151.

Morjan CL, Rieseberg LH (2004) How species evolve collectively: implications of gene flow and selection for the spread of advantageous alleles. *Molecular Ecology*, **13**, 1341-1356.

830

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

835

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

Peisker M (1986) Models of carbon metabolism in C₃-C₄ intermediate plants as applied to the evolution of C₄ photosynthesis. *Plant, Cell & Environment*, **9**, 627-635.

840

Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**, 945-959.

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

845

Rosenberg NA (2004) DISTRUCT: A program for the graphical display of population structure. *Molecular Ecology Notes*, **4**, 137-138.

850

Sage RF (2016) A portrait of the C₄ photosynthetic family on the 50th anniversary of its discovery: species number, evolutionary lineage, and Hall of Fame. *Journal of Experimental Botany*, **67**, 4039-4056.

Sage RF, Christin PA, Edwards EJ (2011) The C₄ plant lineages of planet earth. *Journal of Experimental Botany*, **62**, 3155-3169.

855

Sage RF, Sage TL, Kocacinar F (2012) Photorespiration and the evolution of C₄ photosynthesis. *Annual Review of Plant Biology*, **63**, 19-47.

860

Schmitt T (2007) Molecular biogeography of Europe: Pleistocene cycles and postglacial trends. *Frontiers in Zoology*, **4**, 11.

Smith BS, Brown WV (1973) The Kranz syndrome in the Gramineae as indicated by carbon isotopic ratios. *American Journal of Botany*, **60**, 505-513.

865

Vincens A (1989) Palaeoenvironmental evolution of the North-Tanganyika basin (Zaire, Burundi, Tanzania). *Review of Palaeobotany and Palynology*, **61**, 69-88.

870 von Caemmerer S, Furbank RT (2003) The C₄ pathway: An efficient CO₂ pump. *Photosynthesis Research*, **77**, 191-207.

Williams BP, Johnston IG, Covshoff S, Hibberd JM (2013) Phenotypic landscape inference reveals multiple evolutionary paths to C₄ photosynthesis. *eLife*, **2**, e00961.

875

Data accessibility

All raw reads are available in the short sequence archive under accession number SRP082653. In the NCBI nucleotide database, all newly assembled chloroplast genomes are available under
880 accessions KX752083-KX752090, *ppc* genes under accessions KX788072-KX788087, and *pck* genes under accessions KX788088-KX788109.

Figure and Table captions

885 **Figure 1.** Competing scenarios for the assembly of a complex trait. **A.** The trait is assembled by sequential fixation mutations within each genetic pool. **B.** Mutations that were fixed in isolation are later assembled via secondary gene flow. The species tree is outlined by thick grey branches, and colored branches indicate novel mutations on individual genes. Individual gene trees are drawn in blue and red on the right. In scenario **A** the histories of adaptive mutations correspond to the history of the rest of the genome and all gene trees are concordant, while in scenario **B** the histories of the adaptive mutations differ from that of the rest of the genome, with gene trees that do not match the species tree.

Figure 2. Comparison of nuclear (on the left) and plastid (on the right) topologies (without branch lengths). The putative origin of individuals with mixed genetic background was added using dashed lines. Branches of the nuclear tree are coloured according to clustering analyses (Fig. 3). Photosynthetic types (PT) and presence of laterally-acquired genes (LGT) are indicated by symbols at the tips; purple bar = presence of *pck-IPI-LGT:C*, blue bar = *ppc-IPI-LGT:M*, orange bar = *ppc-IPI-LGT:C*, dark blue bar = *ppc-IPI-LGT:A*. Geographic origin is indicated on the left. Individuals with mixed genetic background were not included in the phylogeny reconstruction, but their putative origin was added using dotted lines. Secondary gene flow are numbered as in Fig. 6; 3) hybridization between non-C₄ and C₄ populations within the Zambezian region, 4) allopolyploidy between C₄ populations in Africa ('RSA3'), 5) pollen-mediated gene flow from mainland Africa to Madagascar.

905

Figure 3. Assignment of *Alloteropsis semialata* individuals to genetic clusters. **A.** Assignment of each individual to the different clusters (*K* 2-4). The photosynthetic type is indicated by symbols next to the names, as in Fig. 2. **B.** Mean likelihood (\pm s.d.) over 10 runs for each *K* value (1-10), and **C.** $|L''(K)|/s.d.$ (fit improvement) as calculated according to Evanno *et al.* (2005).

910

Figure 4. Geographic distribution of *Alloteropsis semialata* genetic lineages. **A.** World distribution, highlighting the Zambezian region with a rectangle and **B.** details of the Zambezian region. For each point, the colour of the outline indicates the plastid lineage (blue = clade A; green = clade BC; yellow = clade FG; red = clade DE), while the colour of the background represents the nuclear lineages (blue = clade I; green = clade II; yellow = clade III; red = clade IV; black = mixed genetic background; grey = congeners). Finally, the shape of the point indicates the photosynthetic type, as determined by carbon isotopes (square = non-C₄; circle = C₄; triangle = isotopically intermediate).

915

Figure 5. Divergence times of C₄ accessions of *A. semialata* based in vertically-inherited and laterally-acquired genes. For five *ppc* and *pck* genes, the posterior distribution of times to the last common ancestor of C₄ *A. semialata* is shown, in million years (Ma).

Figure 6. Inferred history of divergence, secondary exchanges and spread of laterally-acquired *ppc* genes in *A. semialata*. A summary phylogeny is shown for the C₄ and non-C₄ accessions of *A. semialata*, excluding the non-C₄ individuals from South Africa. The C₄ phenotype is represented with red outlines. 1) The divergence of photosynthetic types is inferred in the Zambezian region (dashed red line indicates C₄ emergence). 2) A C₄ lineage migrated outside of the Zambezian region. 3) Hybridization occurred between non-C₄ and C₄ populations within the Zambezian region. 4) The C₄ polyploids of South Africa (RSA) resulted from segmental allopolyploidy. 5) Pollen-mediated gene flow occurred from mainland Africa to Madagascar. The lateral acquisition of three *ppc* genes is indicated with dashed lines, and their subsequent spread is indicated with solid lines. Geographic regions are indicated at the bottom.

Table 1. Results of ABBA-BABA tests.

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

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

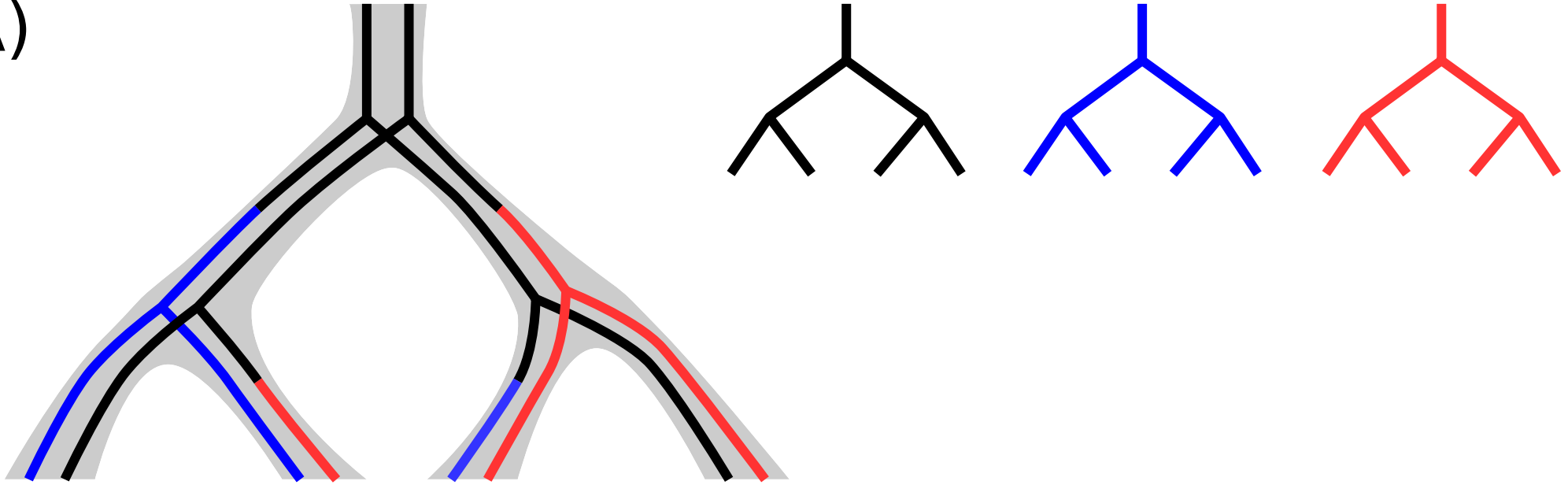
Table 2. Number of read assigned to each of the laterally-acquired *pck* and *ppc* genes.

Species	Accession	Phylogenetic group (plastid; nuclear)	<i>ppc-1P3_LGT:A</i> ¹	<i>ppc-1P3_LGT:M</i> ²	<i>ppc-1P3_LGT:C</i> ³	<i>pck-1P1_LGT:C</i> ³
<i>A. cymicina</i>	Cim1	-	0	149 ⁴	0	0
<i>A. paniculata</i>	Pan1	-	0	37 ⁴	0	0
<i>A. angusta</i>	Ang2	-	0	0	0	78
	Ang1	-	0	0	0	49
<i>A. semialata</i>	RSA1	A; I	0	0	0	0
	RSA2	A; I	0	0	0	0
	TAN1	C; II	0	0	0	57
	TAN2	B; II	0	0	0	73
	TAN3	B; mixed	0	54 ⁴	0 ⁵	216 ⁴
	DRC1	G; mixed	0	57 ⁴	56	183 ⁴
	DRC2	G; mixed	0	29	50 ⁴	95 ⁴
	DRC3	E; III	0	6	25	135 ⁴
	DRC4	E; III	0	10	12	88 ⁴
	TAN4	F; III	0	76	0	83
	RSA3	E; IV	0	55 ⁴	63	113
	KEN1	E; IV	0	36	0	85
	BUR1	E; IV	0	26	0	130
	MAD1	D; IV	0	46	0	101
THA1	D; IV	0	0	0	123	
TPE1	D; IV	0	0	0	118	
AUS1	D; IV		55	0	0	110

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

Fig. 1

A)



B)

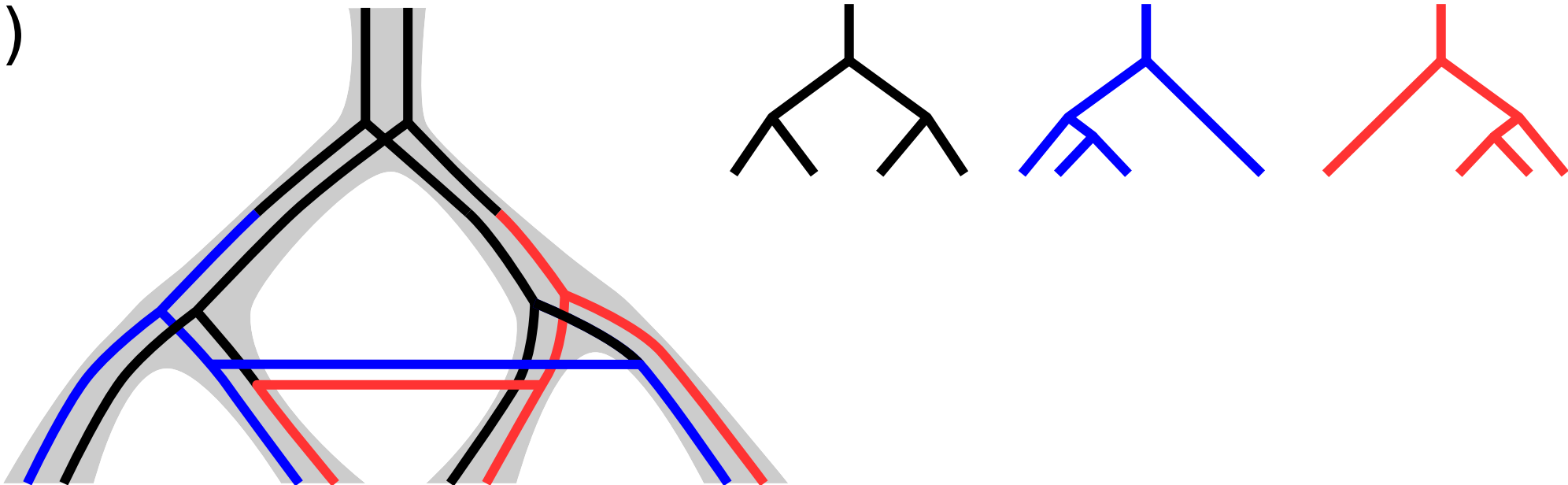


Fig. 2

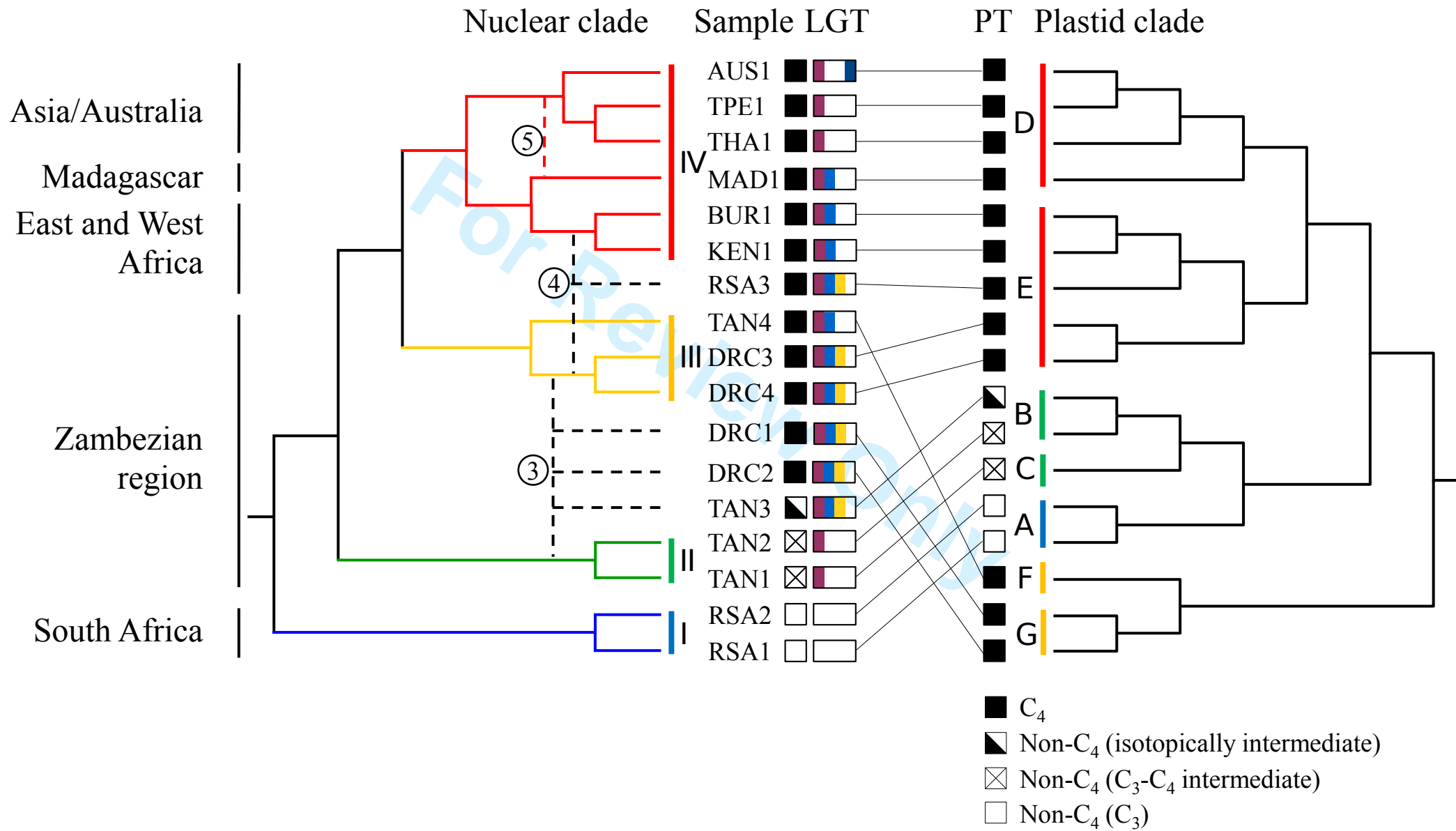


Fig. 3

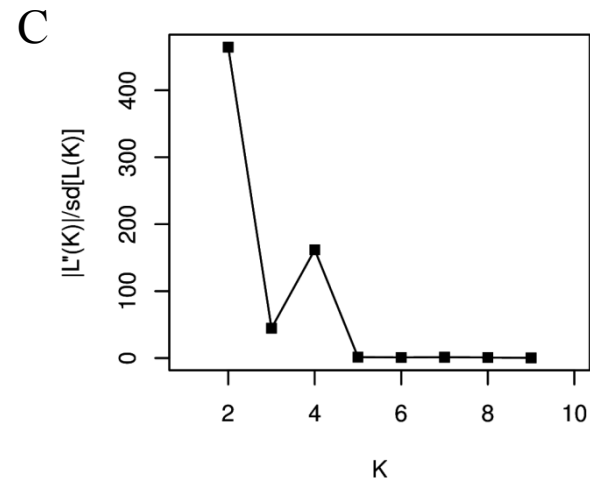
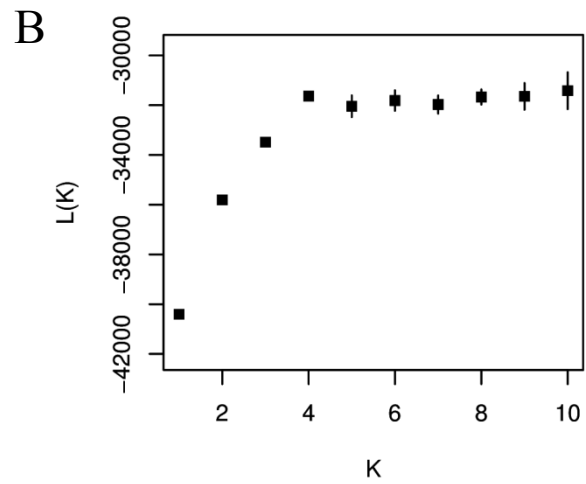
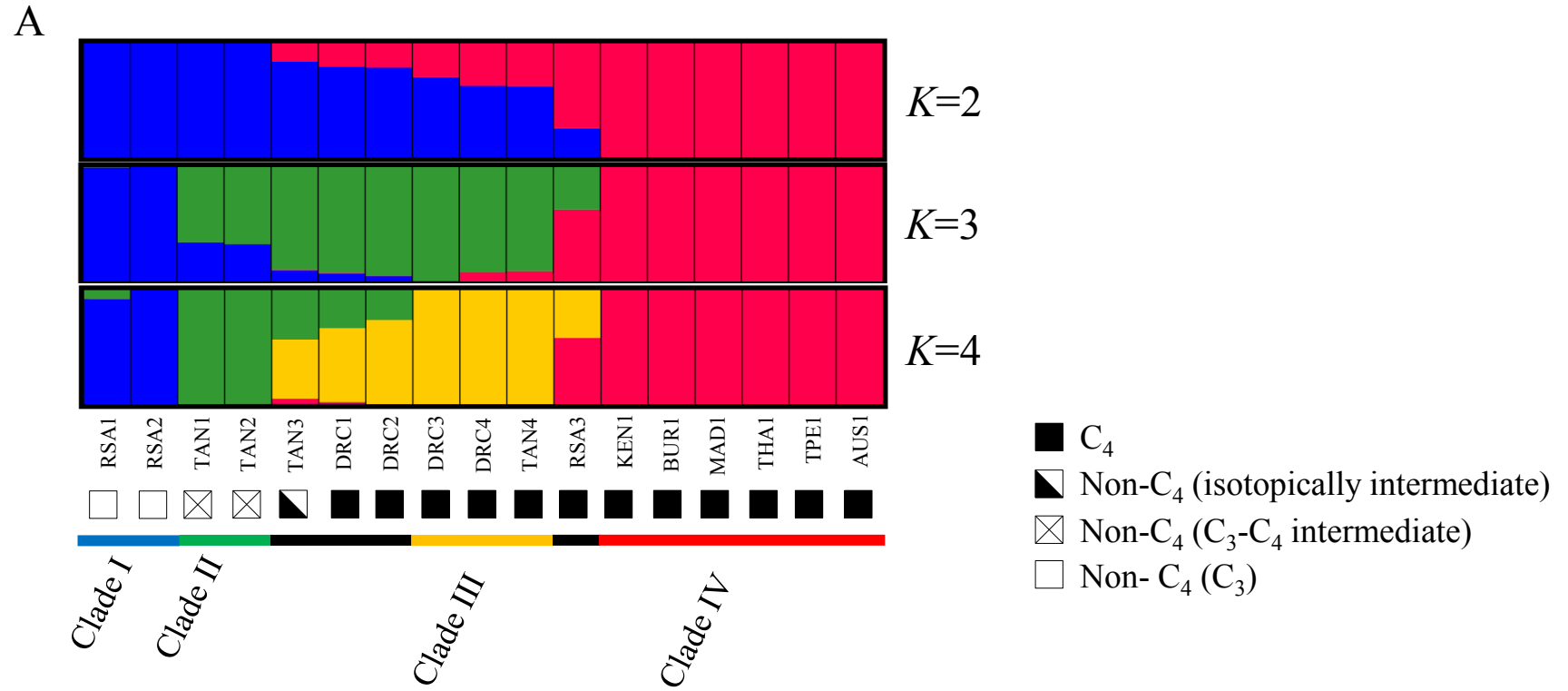
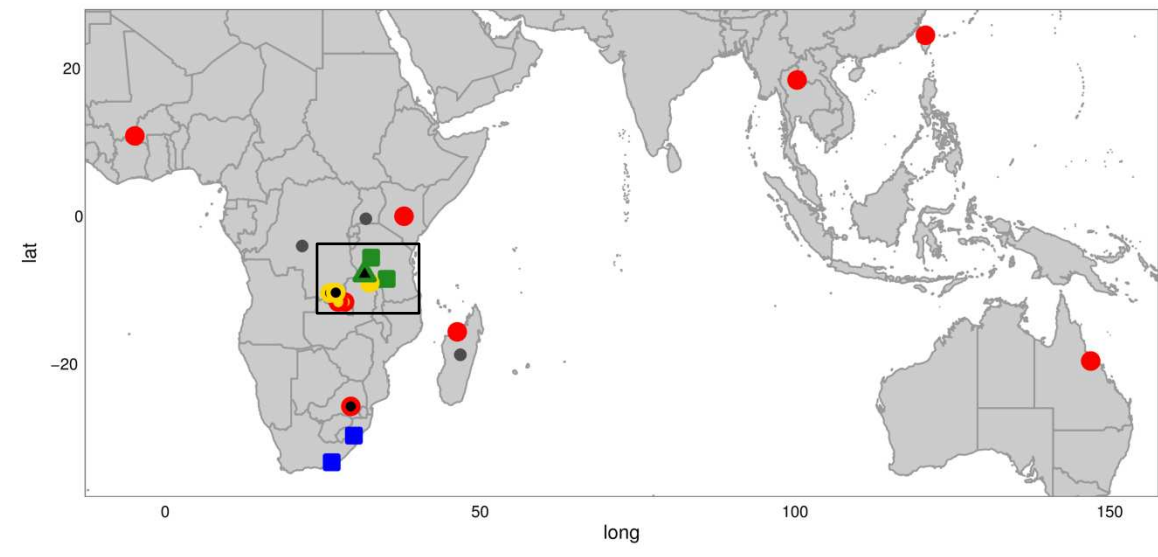


Fig. 4

A



B

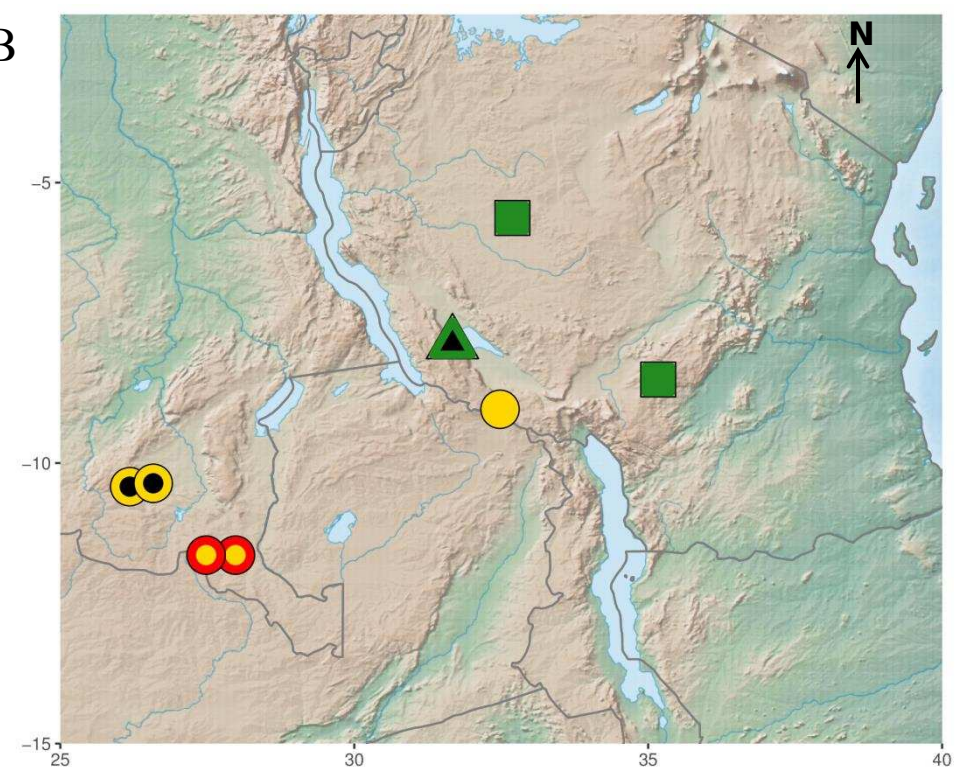
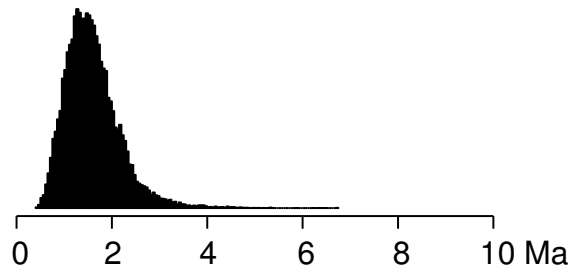
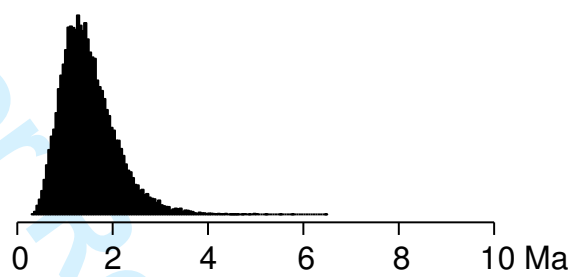


Fig. 5

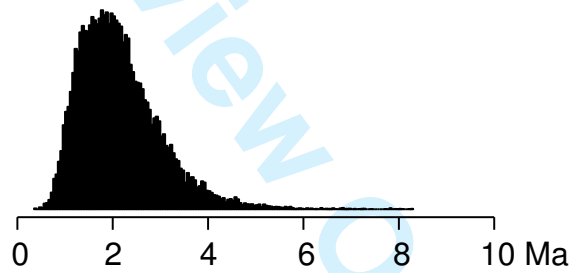
pck-1P1, vertically inherited



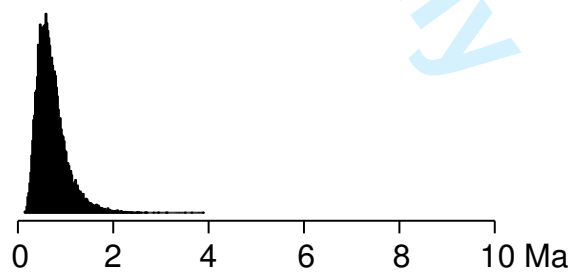
pck-1P1, LGT:C



ppc-1P3, vertically inherited



ppc-1P3, LGT:M



ppc-1P3, LGT:C

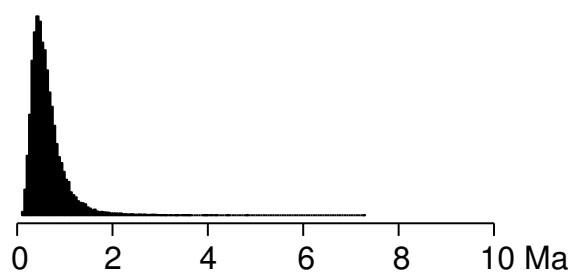


Fig. 6

