Physiological diversity enhanced by recurrent divergence and secondary gene flow within a grass species

Matheus E. Bianconi^{1*}, Luke T. Dunning¹, Emma V. Curran¹, Oriane Hidalgo^{2,3}, Robyn F. Powell², Sahr Mian², Ilia J. Leitch^{2s}, Marjorie R. Lundgren^{1,4}, Sophie Manzi⁵, Maria S. Vorontsova², Guillaume Besnard⁵, Colin P. Osborne¹, Jill K. Olofsson^{1,6}, Pascal-Antoine Christin¹

¹ Department of Animal and Plant Sciences, University of Sheffield, Western Bank, Sheffield S10 2TN, UK; ² Comparative Plant & Fungal Biology, Royal Botanic Gardens, Kew, Richmond, Surrey TW9 3AB, UK; ³ Laboratori de Botànica, Facultat de Farmàcia, Universitat de Barcelona, Unitat Associada CSIC, Av. Joan XXIII s.n., 08028 Barcelona, Catalonia, Spain (current address); ⁴ Lancaster Environment Centre, Lancaster University, Lancaster, LA1 4YQ, UK (current address); ⁵ Laboratoire Evolution & Diversité Biologique (EDB UMR5174), Université de Toulouse III – Paul Sabatier, CNRS, IRD, 118 route de Narbonne, 31062 Toulouse, France; ⁶ Section for GeoGenetics, GLOBE Institute, University of Copenhagen, Øster Farimagsgade 5, building 7, 1353 København K, Denmark (current address)

* author for correspondence: m.bianconi@sheffield.ac.uk, telephone: 01142220027

Short title: Phylogeography of C₄ evolution in a grass species

Summary

- C₄ photosynthesis evolved multiple times independently in angiosperms, but most origins are relatively old so that the early events linked to photosynthetic diversification are blurred. The grass *Alloteropsis semialata* is an exception, as this single species encompasses C₄ and non-C₄ populations.
- Using phylogenomics and population genomics, we infer the history of dispersal and secondary exchanges before, during, and after photosynthetic divergence in *A. semialata*. We further establish the genetic origins of polyploids in this species.
- Organelle phylogenies indicate limited seed dispersal within the Central Zambezian region of Africa, where the species originated ~ 2–3 Ma. Outside this region, the species spread rapidly across the paleotropics to Australia. Comparison of nuclear and organelle phylogenies and analyses of whole genomes reveal extensive secondary gene flow. In particular, the genomic group corresponding to the C₄ trait was swept into seeds from distinct geographic regions. Multiple segmental allopolyploidy events mediated additional secondary genetic exchanges between photosynthetic types.
- Limited dispersal and isolation allowed lineage divergence, while episodic secondary exchanges led to the pollen-mediated, rapid spread of the derived C₄ physiology. Overall, our study suggests that local adaptation followed by recurrent secondary gene flow promoted physiological diversification in this grass species.

Key words: Admixture, C₄ photosynthesis, evolutionary novelty, miombo woodlands, phylogenomics, phylogeography, Poaceae, polyploidy.

Introduction

Most terrestrial plant species assimilate carbon using the ancestral C₃ photosynthetic metabolism, where atmospheric CO₂ is directly fixed by Rubisco in the initial reaction of the Calvin-Benson cycle. The efficiency of this pathway decreases in conditions that limit CO₂ availability, such as warm, arid and saline open habitats (Percy & Ehleringer, 1984; Ehleringer & Monson, 1993; Sage, 2004). In such open habitats of tropical and subtropical regions, C₃ plants are generally replaced by species using the derived C₄ photosynthetic metabolism (Ehleringer & Monson, 1993; Edwards *et al.*, 2010; Osborne *et al.*, 2014). C₄ physiology results from the concerted action of numerous enzymes and anatomical features that together concentrate CO₂ within the leaf and boost productivity in warm, open habitats (Hatch, 1987; Sage, 2004; Atkinson *et al.*, 2016). C₄ plants, and in particular those belonging to the grass and sedge families, nowadays dominate tropical grasslands and savannas, which they have shaped via feedbacks with herbivores and fire (Scheiter *et al.*, 2012; Osborne *et al.*, 2014; Sage & Stata, 2015; Lehmann *et al.*, 2019). The C₄ metabolism evolved multiple times independently and its origins are spread over the past 30 million years (Christin *et al.*, 2008, 2011; Sage *et al.*, 2011). Retracing the eco-evolutionary dynamics linked to photosynthetic transitions is difficult for old C₄ lineages, but a few lineages evolved the C₄ trait relatively recently.

The grass *Alloteropsis semialata* is the only species known to have genotypes with distinct photosynthetic pathways (Ellis, 1974, 1981). C₄ accessions are distributed across the paleotropics to Oceania, while C₃ individuals are restricted to southern Africa (Lundgren *et al.*, 2016; Fig. 1). In addition, individuals performing a weak C₄ pathway (C₃+C₄ individuals *sensu* Dunning *et al.*, 2017) occur in parts of Tanzania and Zambia (Lundgren *et al.*, 2016). Their distribution matches the plant biogeographical region referred to as ‘Zambezian’ (Linder *et al.*, 2012) or ‘Central Zambezian’ (Droissart *et al.*, 2018) and are associated with miombo woodlands (Burgess *et al.*, 2004). Plastid genome analyses have suggested that the species originated in this region, where divergent haplotypes are still found (Lundgren *et al.*, 2015; Olofsson *et al.*, 2016). One lineage associated with the C₃ type then migrated to southern Africa, while a C₄ lineage dispersed across Africa and Asia/Oceania (Lundgren *et al.*, 2015). The species phylogeography was previously studied using complete plastid genomes from just a few individuals coupled with individual nuclear markers from a denser sampling, leading to limited resolution (Lundgren *et al.*, 2015). The exact origin of the different lineages thus remains to be firmly established.

Whether the different physiological types of *A. semialata* belong to the same species was originally questioned (Gibbs Russell, 1983), as they were shown to be associated with distinct ploidy levels in South Africa, where the C₃ are diploid and the C₄ vary from tetra- to dodecaploids (Ellis, 1981; Liebenberg & Fossey, 2001). However, C₄ diploids have been reported in eastern and

western Africa, Madagascar, Asia and Australia (Ellis, 1981; Lundgren *et al.*, 2015), and nuclear genome analyses have found evidence of genetic exchanges between lineages with different photosynthetic types (Olofsson *et al.*, 2016). In addition, discrepancies between mitochondrial and plastid genomes (Olofsson *et al.*, 2019b) might reflect the footprint of segmental allopolyploidy. However, the history of nuclear exchanges and their effect on the spread of different photosynthetic types through ecological and geographic spaces remains to be formally established.

In this study, we analyse the organelle and nuclear genomes of 69 accessions of *A. semialata* from 28 countries, covering the known species range and different photosynthetic types, to establish the order of seed-mediated range expansion and subsequent pollen-mediated admixture of nuclear genomes. Organelle phylogenetic trees are used to (1) identify the geographic and ecological origins of the species and its subgroups, with a special focus on the C₄ lineages. Analyses of nuclear genomes are then used to (2) establish the history of secondary genetic exchanges and their impact on the sorting of photosynthetic types. Finally, genome size estimates coupled with phylogenomics and population genomics approaches are used to (3) infer the origins of the polyploids and their relationship to photosynthetic divergence. Our detailed genome biogeography analyses shed new light on the historical factors that lead to functional diversity within a single species.

Materials and Methods

Sampling, sequencing and data filtering

Whole genome sequencing data from 69 accessions of *Alloteropsis semialata* (R. Br.) Hitchc. were used in this study, along with two samples of each of its congeners *A. cimbicina* (L.) Stapf, *A. paniculata* (Benth.) Stapf, and *A. angusta* Stapf. Genomic datasets of 48 accessions were retrieved from previous studies (Lundgren *et al.*, 2015; Olofsson *et al.*, 2016, 2019b; Dunning *et al.*, 2019b; Table S1). Another 26 accessions of *A. semialata* and one of *A. paniculata* were sequenced here (Table S1). For the new samples, genomic DNA (gDNA) was isolated from herbarium samples using the BioSprint 15 DNA Plant Kit (Qiagen), or from fresh or silica gel dried leaves using the Plant DNeasy Extraction Kit (Qiagen; Table S1). Libraries from herbarium samples were prepared with 22–157 ng of gDNA using Illumina TruSeq Nano DNA LT Sample Prep kit (Illumina, San Diego, CA, USA). They were sequenced at the GenoToul-GeT-PlaGE platform (Toulouse, France), as paired-end reads on 1/24th of an Illumina HiSeq3000 lane (Table S1). For fresh or silica gel dried samples, libraries were constructed by the respective sequencing facilities, and paired-end sequenced on full, 1/6th or 1/12th lane of an Illumina HiSeq2500 at the Sheffield Diagnostic Genetics Service (UK) and the Edinburgh Genomics facility (UK; Table S1). Raw Illumina datasets were filtered before analysis using the NGSQC Toolkit v.2.3.3 (Patel & Jain, 2012) to remove low

quality reads (i.e. < 80% of the bases with Phred quality score > 20), reads containing ambiguous bases and adaptor contamination. The retained reads were further trimmed from the 3' end to remove bases with Phred score < 20. The quality of the filtered datasets was assessed using FastQC v0.11.9 (Andrews, 2010).

Genome sizing and carbon isotope analyses

Genome sizes of *A. semialata* accessions were either retrieved from previous studies (Lundgren *et al.*, 2015; Olofsson *et al.*, 2016) or estimated here from fresh or silica gel dried leaves using flow cytometry. For these, we used the one-step protocol of Doležel *et al.* (2007) with minor modifications (Clark *et al.*, 2016). We selected as internal standards either *Petroselinum crispum* 'Champion Moss Curled' (2C = 4.5 pg; Obermeyer *et al.*, 2002) or *Oryza sativa* IR36 (2C = 1.0 pg; Bennett & Smith, 1991) for diploid accessions, and *Pisum sativum* 'Ctirad' (2C = 9.09 pg; Doležel *et al.*, 1992) for accessions whose C-value was estimated to be more than three times larger than that of diploid accessions. For fresh leaf material either the Ebihara (Ebihara *et al.*, 2005) or the GPB buffer (Loureiro *et al.*, 2007) supplemented with 3% of PVP was used as the nuclei isolation buffer, whereas when only silica-dried material was available the Sysmex CyStain PI Oxprotect isolation buffer was used. The prepared material was analysed using a Sysmex Partec Cyflow SL3 flow cytometer (Sysmex, Germany) fitted with a 100-mW green solid-state laser (Cobalt Samba, Solna, Sweden). Using genome sizes previously estimated from individuals with a chromosome count (Lundgren *et al.*, 2015) enabled the ploidy level to be assigned to all individuals with a genome size estimated here.

Carbon isotope composition was used to determine the photosynthetic type of each accession. These data were either retrieved from previous studies (Lundgren *et al.*, 2015, 2016, 2019; Olofsson *et al.*, 2016) or measured here (Table S1). Dry leaves were ground to a fine powder using a TissueLyzer (Qiagen) and 1-2 mg of material was analysed using an ANCA GSL preparation module coupled to a Sercon 20-20 stable isotope ratio mass spectrometer (PDZ Europa, Cheshire, UK). Carbon isotopic ratios ($\delta^{13}\text{C}$, in ‰) were reported relative to the standard Pee Dee Belemnite (PDB). In C_4 plants, $\delta^{13}\text{C}$ values range between -16 and -10‰ (Smith & Brown, 1973). All individuals with $\delta^{13}\text{C}$ values below -16‰ were classified as non- C_4 . These were further distinguished between C_3 and C_3+C_4 using previous anatomical and physiological data, and/or expression levels of C_4 pathway genes, where available (Dunning *et al.*, 2019a; Lundgren *et al.*, 2016, 2019).

Assembly of organelle genomes and molecular dating

Complete mitochondrial and plastid genomes were retrieved from previous studies (Lundgren *et al.*, 2015; Olofsson *et al.*, 2016) or assembled here using a reference-guided approach. As a reference dataset, we used the chromosome-level nuclear, mitochondrial and plastid genomes of an Australian individual of *A. semialata* (Lundgren *et al.*, 2015; Dunning *et al.*, 2019b). Paired-end genomic reads were mapped to the reference using Bowtie2 v2.3.5 (Langmead & Salzberg, 2012) with default parameters. Reads uniquely mapped to the plastid genome were extracted and re-mapped as pairs to the plastid reference sequence using Geneious v6.1.8 (Kearse *et al.*, 2012) with medium sensitivity and three iterations for fine tuning. The final plastome assembly consisted of the majority consensus sequence from the mapped reads. Plastome sequences were aligned using MAFFT v7.427 (Katoh & Standley, 2013), and further curated after visual inspection using the block realignment option in Aliview v1.17.1 (Larsson, 2014). The second inverted repeat was removed, resulting in a 121,059 bp alignment. The same approach resulted in a low-quality alignment when applied to the mitochondrial genome, possibly because genome rearrangements created large discontinuities in the assembled sequences. To obtain mitochondrial genomes from the initial read alignment files, variant sites were extracted from the reads uniquely mapped to the mitochondrial genome and incorporated into a consensus sequence using the mpileup function of Samtools v1.9 (Li *et al.*, 2009) implemented in a bash-scripted pipeline (Olofsson *et al.*, 2019a). Only reads with mapping quality > 20 and bases covered by more than 50% of the reads were retained, and sites covered by less than ten reads were called as unknown bases. This approach provides sequences that are already aligned. The alignment was then trimmed to remove sites with more than 10% missing data using trimAl v1.4 (Capella-Gutiérrez *et al.*, 2009), resulting in a 136,410 bp alignment.

A time-calibrated phylogeny was obtained independently on plastid and mitochondrial alignments using BEAST v1.8.4 (Drummond & Rambaut, 2007). The median ages estimated by Lundgren *et al.* (2015) were used for secondary calibration of the genus *Alloteropsis* (11.46 Ma for the root, and 8.075 Ma for the split between *A. angusta* and *A. semialata*), both using a normal distribution with standard deviation of 0.0001. Analyses were performed using the GTR+G+I model of substitution and a lognormal uncorrelated relaxed clock (Drummond *et al.*, 2006), with a constant population size coalescent tree prior. Two MCMC analyses were run in parallel for 300,000,000 generations with sampling every 20,000 generations using the CIPRES Science Gateway v3.3 (Miller *et al.*, 2010). Convergence of runs and effective sample sizes > 100 were confirmed by monitoring the log files using Tracer v1.6 (Rambaut *et al.*, 2013). The burn-in period was set after the convergence of the runs at 10% of generations, and median ages of posterior trees were mapped on the maximum clade credibility tree.

Phylogenetic analyses of the nuclear genome

Genome-wide nuclear markers were assembled using the genomic data and combined into a multigene coalescent phylogeny. Putative single-copy orthologs of Panicoideae were first identified from the genomes of *A. semialata* (Dunning *et al.*, 2019b), *Setaria italica*, *Panicum hallii* and *Sorghum bicolor* (the last three extracted from Phytozome v13; Goodstein *et al.*, 2012) using OrthoFinder v2.3.3 (Emms & Kelly, 2019). A total of 7,408 genes were identified as single-copy orthologs and *A. semialata* coding sequences (CDS) were extracted and used as the reference. Sequences for each reference CDS were then assembled from all *Alloteropsis* accessions using the approach described above for mitochondrial genomes, except that reads were mapped as unpaired to avoid discordant pairs where mates mapped to non-exonic sequences. Gene alignments were trimmed using trimAl to remove sites covered by less than 70% of individuals (more than 30% missing data), and individual sequences shorter than 200 bp after trimming were discarded. Only trimmed gene alignments longer than 500 bp and with more than 95% of taxon occupancy were retained. Maximum-likelihood phylogenies were inferred for each of these 3,553 alignments using RAxML v8.2.4 (Stamatakis, 2014), with a GTR+CAT substitution model and 100 bootstrap pseudoreplicates. The gene trees were summarized into a multigene coalescent phylogeny using Astral v5.6.2 (Zhang *et al.*, 2018) after collapsing branches with bootstrap support values below 30. To verify that the nuclear analyses were not distorted by the presence of polyploids, we estimated an additional multigene coalescent phylogeny using only confirmed diploid samples. To evaluate the effect of less stringent missing data thresholds on the resulting tree, we also repeated the Astral analysis using the filtering strategy described above, except that we used alignments that (1) were not trimmed, and (2) were trimmed to remove sites covered by less than 30% of individuals (more than 70% missing data). In both cases, short alignments and those with a low taxon occupancy were removed as before.

Genetic structure

To examine the genetic structure of *A. semialata* across its known range, a principal component analysis and an individual-based admixture analysis were performed. These analyses used the read alignments corresponding to the whole nuclear genome in the original mapping described above for organelle genome assemblies. Mapped reads were sorted and indexed using Samtools v1.9, and duplicate reads were removed using the function MarkDuplicates from Picard tools v2.13.2 (<http://broadinstitute.github.io/picard/>). Genotype likelihoods were estimated using ANGSD v0.929 (Korneliussen *et al.*, 2014). Sites covered by at least 70% of individuals and with a minimum mapping and base quality score of 30 were retained, resulting in a set of 11,439 variable sites. A minimum depth of coverage per site per individual was not specified, to incorporate some individuals with a low average sequencing depth ($< 0.5 \times$; Table S1). A covariance matrix was

estimated from the genotype likelihoods using PCAngsd v0.98 (Meisner & Albrechtsen, 2018). Eigenvector decomposition was carried out using the *eigen* function in R version 3.4.4 (R Core Team, 2018) to recover the principal components of genetic variation. Individual admixture proportions were estimated from genotype likelihoods using a maximum likelihood approach with NGSadmix v32 (Skotte *et al.*, 2013). NGSadmix was run with different numbers of ancestral populations (*K*) ranging from 1 to 10, with five replicates for each value, each with a different random starting seed. The value of *K* that best describes the uppermost level of structure in the data was estimated using the Evanno method (Evanno *et al.*, 2005), as implemented in CLUMPAK (Kopelman *et al.*, 2015).

Genome composition

A pipeline was developed to quantify the proportion of alleles in each genome that were descended from each of the main four nuclear clades of *A. semialata* (see Results). A reference dataset was first built with putative single-copy genes in land plants (Embryophyta) obtained from BUSCO v3.0.2 (Simão *et al.*, 2015), using, for each group of orthologs, the sequence extracted from the reference genome of *A. semialata*. Out of 1,202 BUSCO genes identified in *A. semialata*, a total of 473 had at least one exon longer than 550 bp (which corresponds to the average insert size of the paired-end genomic datasets used here). The longest exon of each of these genes was used for subsequent analyses.

We used data for 26 individuals of *A. semialata* that were diploid, including a F1 hybrid between C₃ (nuclear clade I) and C₄ (nuclear clade IV) parents (Table S2). Four congeners (three *A. angusta* and two *A. cimicina*) were added to serve as outgroups. All 30 individuals were sequenced as 250-bp paired-end reads (insert size = 550 bp) and the data were mapped to the reference genome of *A. semialata* (Dunning *et al.*, 2019b) using Bowtie2 with default parameters, except the insert size (-X) that was increased to 1,100 bp. Mapped reads were then phased for the 473 exons using the phase function of Samtools v0.1.19, where bases with quality < 20 were removed during heterozygous calling (option -Q20), and reads with ambiguous phasing were discarded (option -A). Sequences were then generated for both alleles using custom bash scripts in which different depth filters were applied for resequencing and high coverage ($\geq 20 \times$) datasets (≥ 3 and ≥ 10 reads covering each position, and ≥ 2 and ≥ 3 reads covering each variant for polymorphic sites, respectively). All phased sequences shorter than 200 bp were discarded. To select only alignments with sufficient information, we only kept genes that: (i) included at least one allele for each of the two congeners *A. cimicina* and *A. angusta*; (ii) had at least four alleles for each of the four nuclear clades of *A. semialata*; and (iii) had at least 30 alleles in total (50% of the possible maximum). A maximum likelihood tree was inferred for each of the 300 genes that matched these criteria using

PhyML v20120412 (Guindon *et al.*, 2010) with a GTR+G+I model. The resulting trees were rooted on *A. ciminica* and processed with custom scripts to identify those genes for which one allele of the F1 hybrid was nested within the C₃ clade I and one nested within the C₄ clade IV. The 180 markers fulfilling this criterion were deemed phylogenetically informative and used for final analyses.

For five hexa- and dodecaploids of *A. semialata* sequenced as 250 paired-end reads, individual alignments were successively generated for each pair of reads that fully overlapped with one of the 180 exons, as determined using BEDTools v2.24.0 (Quinlan & Hall, 2010), with the function 'intersect' (option -f 1.0). Each read pair was separately added to the respective gene alignment using the MAFFT function '-add'. The paired reads were then merged, and each alignment was trimmed to remove non-overlapping regions (max. alignment length = 500 bp). Samples with > 50% missing data in this truncated alignment (i.e. < 250 bp) were subsequently discarded. A maximum likelihood tree was then inferred as described above, and the merged read pair was assigned to the nuclear clade in which it was nested. Read assignments were not used in cases in which the C₃ and C₄ alleles from the F1 hybrid were not correctly placed as a consequence of alignment trimming, or when the sister group of the reads was composed of multiple lineages. The analyses were later repeated with each diploid used as the focus species, in which case the phased alleles from the focal individual were removed from the reference dataset.

Results

Genome sizes

Out of 38 samples for which genome sizes are available, 31 ranged from 1.78 to 2.77 pg/2C, which are values $\pm$ 25% within the range previously reported for diploid *A. semialata* ($2n = 2x = 18$) (Olofsson *et al.*, 2016; Table S1). One individual from Australia has a genome size of 3.65 pg/2C, which suggests tetraploidy ($2n = 4x = 36$), while three individuals from Zambia and one from Mozambique have genome sizes between 5.35 and 6.71 pg/2C, which suggests hexaploidy ($2n = 6x = 54$), as previously reported in South Africa (Liebenberg & Fossey, 2001; Lundgren *et al.*, 2015). Finally, individuals from one Cameroonian population had genome sizes suggesting dodecaploids ($2n = 12x = 108$; 11.87 pg/2C; Table S1). All polyploids detected in *A. semialata* so far have isotopic signatures of C₄ plants (Table S1).

Time-calibrated organelle phylogenies

The phylogenetic trees based on plastid and mitochondrial genomes recovered the seven major lineages reported in previous studies (Figs 2, S1; Lundgren *et al.*, 2015; Olofsson *et al.*, 2016, 2019b). The relationships were mostly congruent among the two organelles, although some samples were positioned in different groups as previously reported (Olofsson *et al.*, 2019b). Indeed, six

samples from Cameroon, Democratic Republic of Congo (DRC), and Zambia form a clade within the DE lineage in the plastid phylogeny, but form a paraphyletic group within the FG mitochondrial lineage (Fig. S1; Olofsson *et al.*, 2019b). The genome sizes that are available for samples from this group indicate they are either hexa- or dodecaploids (Table S1).

The first split within *A. semialata*, estimated at 2.9/2.1 Ma on the mitochondrial/plastid trees (95% HPD = 1.6 – 4.5 / 1.4 – 3), separates a lineage FG occurring in the Central Zambezan region of Africa (in DRC, Zambia and Tanzania) from accessions spread around the world (ABCDE). Within the FG group, accessions from DRC diverge first, and accessions from the south of Tanzania are nested within a paraphyletic Zambian clade (Figs 2, S1). Within the ABCDE group, the first split (2.1/1.9 Ma) separates the non- C_4 group ABC from the exclusively C_4 lineage DE (Lundgren *et al.*, 2015). Accessions from the B and C clades are spread across central/northern areas of the Central Zambezan region (Burundi, DRC, Tanzania and Zambia), while clade A is composed of accessions from southern Africa, with early divergence from accessions from Mozambique and Zimbabwe that likely represents the footprint of a gradual migration to South Africa between 0.8 and 0.4 Ma (Figs 2, S1).

Within the large C_4 clade, lineage D contains accessions from Asia and Oceania that are sister to samples from Madagascar, as previously reported (Lundgren *et al.*, 2015). Together, these populations are sister to a sample from Angola (Figs 2, S1). The sister lineage E contains a subgroup spread east of the Central Zambezan region (Tanzania, Malawi and Mozambique) and South Africa, while the other subgroup contains one accession from Ethiopia that is sister to samples spread from Kenya to Sierra Leone with very little divergence (Figs 2, S1).

Nuclear phylogeny

A multigene coalescent phylogeny was obtained for all 75 accessions of *Alloteropsis*, using 3,553 nuclear markers. The monophyly of *A. semialata* was supported with little evidence of incomplete lineage sorting (Fig. 3). Within *A. semialata*, the four major nuclear clades were retrieved with high support (Fig. 3; Olofsson *et al.*, 2016). Similar relationships were obtained when different thresholds for missing data were used (Fig. S2), and when only diploid samples were included (Fig. S3). The lineages identified based on organellar genomes were mostly recovered by the nuclear analyses, although the relationships among them differed, as previously reported (Olofsson *et al.*, 2016; Dunning *et al.*, 2019b). Nuclear clade I corresponds to organelle lineage A, which is associated with C_3 photosynthesis. Nuclear clade II contains non- C_4 accessions from the Central Zambezan region (C_3+C_4 ; organelle lineages B and C), and it is placed as sister to clades III+IV (Fig. 3) or to clade II (Figs S2, S3), in both cases with a similar number of quartets supporting the alternative topology.

Among C_4 accessions, one Tanzanian and two Congolese accessions (from organelle lineages B and F, respectively) are placed at the base of nuclear clades III and IV with low quartet support values (Figs 3, S2). These three accessions were previously reported as admixed individuals with nuclear contributions from clades II and III (Olofsson *et al.*, 2016). Two accessions from Cameroon, one of which is a known dodecaploid, are in a similar position (Fig. 3). These five accessions are subsequently referred to as the 'unplaced' nuclear group. Nuclear clade III is formed exclusively of C_4 accessions from the mitochondrial lineage FG (including those placed within plastid lineage E) that are restricted to the Central Zambebian region. Nuclear clade IV contains all C_4 individuals from mitochondrial lineage DE, but with relationships that differ from the organelle genomes. The first splits lead to South African hexaploids, one Mozambican hexaploid accession and one Angolan accession, while most accessions cluster in one of two sister groups; one composed of all other African accessions (including Madagascar) and one composed exclusively of Asian and Oceanian accessions (Figs 3, S2, S3).

Population structure and genome composition

A principal component analysis grouped individuals largely according to their nuclear phylogenetic relationships (Fig. S4a, b). Admixture analyses identified four and seven clusters as good-fit for the data (Fig. S4c), and again retrieved groups that match the nuclear phylogeny (Fig. 4a). In particular, the five unplaced individuals were also positioned in between clades in the principal component analysis (Fig. S4a) and showed mixed ancestry (Fig. 4a).

Phylogenetic analyses of exons of putative single-copy genes were used to further evaluate the genomic composition of a subset of accessions for which longer paired-end reads were available. As expected, more than 90% of the reads from the Asian/Oceanian individuals were assigned to the C_4 nuclear clade IV (Olofsson *et al.*, 2016; Fig. 4b; Table S2). Low levels of assignment to other nuclear clades might represent incomplete lineage sorting or methodological noise. The majority of reads from the African diploids were similarly assigned to the expected nuclear clade, but up to 18% of the reads from these individuals were assigned to other clades. In the C_3 clade I, more than 90% of the reads from South African samples were assigned to clade I, but this number dropped to 78% in the sample from Zimbabwe. In this latter sample, which is geographically closer to the Central Zambebian region (Fig. 2), 15% of the reads were assigned to clade II. The proportion of reads of individuals from clade II assigned to the expected clade varied between 82 and 93%, but the proportion of reads of C_4 diploids from Africa assigned to the expected clade was as low as 68%, with up to 18% of reads assigned to the other C_4 clade and up to 11% assigned to the non- C_4 clade II. The ancestry of the polyploid individuals differed between geographic regions. The Mozambican hexaploid was mostly assigned to the C_4 clade IV (80% of

reads), with 15% of the reads assigned to the other C₄ group (clade III). For the South African hexaploids, about 80% of the reads were assigned to the two C₄ clades: ~ 60% to clade IV and ~ 20% to clade III (Fig. 4b; Table S2). Similarly, approximately 80% of the reads from one hexaploid from Zambia were assigned to C₄ clades III and IV, but the proportions were inverted compared to South Africa, with 63% of reads assigned to clade III and 19% to clade IV. Finally, the reads from the dodecaploid individual from Cameroon are almost equally spread among the C₄ clades III and IV and the non-C₄ clade II (Fig. 4b; Table S2).

Discussion

Limited seed dispersal in the region of origin

As previously reported (Lundgren *et al.*, 2015), *Alloteropsis semialata* likely originated in the Central Zambezan region of Africa. Four organelle lineages capturing the earliest splits within *A. semialata* (B, C, F and G) are restricted to this region (Figs 2, S1). The geographic distribution of these haplotypes closely matches that of Central Zambezan miombo woodlands (Fig. 2), which are characterized by a more or less continuous grass coverage and the dominance of a few tree genera, particularly *Brachystegia* and *Julbernardia* (Fig. 1; Lawton, 1978; White, 1983; Frost, 1996; Burgess *et al.*, 2004). Palynological evidence indicates that *Brachystegia* was already present in eastern Africa during the late Oligocene (Vincens *et al.*, 2006), and throughout the Miocene (Yemane *et al.*, 1987), which suggests that the first divergence within *A. semialata* (~ 2–3 Ma) happened in this biome. The extent of miombo woodlands varied with glaciation cycles (Beuning *et al.*, 2011; Ivory & Russell, 2016), and restrictions to dispersal during glacial maxima might have driven the vicariance of organelle lineages FG and ABCDE of *A. semialata* in the Pleistocene, as previously reported for other taxa occurring across the Great Rift System (e.g. Mairal *et al.*, 2017). Indeed, the order of splits within group FG is compatible with an origin west of the Rift Valley lakes, while lineages B and C likely originated to the east of the lakes (Fig. 2). The present-day co-occurrence of FG and BC organelle groups probably follows a migration beyond their refugia after the re-expansion of miombo woodlands (Fig. 2; Vincens, 1991; Dupont *et al.*, 2008; Ivory & Russel, 2016). This region of Africa was also a major Pleistocene refugium for several savanna mammals (Lorenzen *et al.*, 2012; McDonough *et al.*, 2015; Pedersen *et al.*, 2018), confirming the broad biogeographical importance of the region (Linder *et al.*, 2012).

Despite having originated about 2 Ma, lineages FG and BC still occur within a relatively small geographic region in central/eastern Africa. The visible geographic structure of each of these lineages in the organelle phylogenies further supports limited seed dispersal, as previously reported (Lundgren *et al.*, 2015). These two lineages occur in the wet miombo that occupies the mountains separating the Zambezi and Congo basins. Variations in elevation coupled with relatively dense tree

cover might limit seed dispersal for this species with seeds that probably spread mainly by gravity. By contrast, the lineages that escaped this centre of origin bear the footprint of a rapid geographic spread. Over the last million years, lineage A migrated to the south of Africa, where cooler climates might have selected for the C₃ photosynthetic type (Long, 1983; Lundgren *et al.*, 2015, 2016). The ancestor of lineage DE likely first reached the lowland surrounding the wet Central Zambezian miombo around 1 Ma (Fig. 2). This lineage then rapidly spread around the world, with part of lineage E migrating to southern Africa, while the rest spread north and rapidly reached western Africa, probably via the Sudanian savanna (Fig. 2; Linder *et al.*, 2012). In the meantime, lineage D spread west to Angola, reached Madagascar in the east, and colonized southeastern Asia and Australia (Lundgren *et al.*, 2015; Olofsson *et al.*, 2019b).

The rapid migration of lineage DE was facilitated by the broader niche conferred by the C₄ photosynthetic type that characterizes this lineage (Lundgren *et al.*, 2015). However, corridors of low elevation, coupled with open grasslands in southern Africa, also likely contributed to the long distance spread of *A. semialata* seeds outside the Central Zambezian region, including the C₃ lineage A.

Widespread pollen flow and sweep of the C₄ nuclear genome

The organelle lineages are loosely associated with distinct nuclear groups. In particular, the strongly supported nuclear clade I (Fig. 3) is exclusively associated with organelle lineage A and encompasses all C₃ individuals from southern Africa (Figs 2 and 3). Most non-C₄ individuals from the Central Zambezian region (organelle lineages B and C) belong to nuclear clade II, while the C₄ organelle group FG mostly matches nuclear clade III, and the C₄ organelle group DE mostly corresponds to nuclear clade IV (Figs 2, 3 and 5). These matches indicate that the split of seed-transported organelles was accompanied by a reduction of nuclear exchanges. Geographic segregation of lineages into distinct regions might have over time led to a partial post-zygotic reproductive isolation in addition to their vicariance. However, the nuclear structure is less clear and numerous discrepancies between nuclear and organelle genomes indicate secondary genetic exchanges mediated by pollen. Such cytoplasmic-nuclear discordances are widespread in plants and animals (Rieseberg & Soltis, 1991; Toews & Brelsford, 2012), and have revealed complex patterns of lineage diversification (e.g. Folk *et al.*, 2017; Lee-Yaw *et al.*, 2019; Muñoz-Rodríguez *et al.*, 2018).

The organelle phylogenetic trees consistently identify two distinct C₄ groups (FG and DE; Figs 2, 3, S1), while all C₄ accessions are monophyletic in the nuclear analyses (Fig. 3; Olofsson *et al.*, 2016; Dunning *et al.*, 2019b). The sister group relationship between nuclear clades III and IV, which are associated with divergent organelles, suggests the swamping of one nuclear genome

lineage by the other (Fig. 5). The directionality of this exchange is unknown, but repeated, unidirectional gene flow mediated by pollen must have occurred from one group into seeds from a different region, as previously reported for other taxa (e.g. Buggs & Pannell, 2006). Based on organelle phylogenies, one of the lineages originates from the Central Zambesian highlands (lineage FG), while the other originates from the lowlands of eastern Africa (lineage DE; Fig. 2). One possibility is that differences in elevation along with predominantly easterly winds could have restricted seed migration but favoured pollen flow from the lowlands to the west, leading to gene flow from lineage DE to FG. Given that C_4 genes are encoded by the nuclear genome, it is tempting to hypothesize that an efficient C_4 trait evolved following migration to lower elevation, where higher temperatures increased selective pressures for photosynthetic transitions (Ehleringer & Monson, 1993; Edwards & Still, 2008). Pollen flow would then have brought the C_4 pathway to populations from the highlands, where selection for the derived pathway would have mediated the sweep. Such a scenario would be compatible with a C_4 origin promoted in a small range of very warm habitats, while the C_4 pathway then broadens the niche and facilitates subsequent transitions to colder environments (Lundgren *et al.*, 2015; Aagesen *et al.*, 2016; Watcharamongkol *et al.*, 2018). Independently of the direction, this marked incongruence between organelle and nuclear phylogenies indicates that the C_4 trait was rapidly spread by hijacking seeds of the same species, in addition to speeding up the dispersal of lineage DE.

Recurrent hybridization and polyploidization

Besides the sweep of the nuclear genome into divergent seeds, the comparison of nuclear genomes suggests episodic hybridization between the different lineages. First, the sister group relationship between the non- C_4 nuclear clades II and I is supported by 40–42% of quartet trees across the different multigene coalescent analyses (Figs. 3, S2, S3), which is closer to the organelle phylogenetic relationships (Figs 2, S1), but a similar proportion of quartets (34–42%) places nuclear clade II as sister to nuclear clades III+IV. Dominance of two alternative topologies over a third one is not predicted under pure incomplete lineage sorting (Sayyari & Mirarab, 2016), and the incongruence between a proportion of the nuclear genes and organelles likely results from an episode of hybridization. It is possible that this episode brought some genes adapted for the C_4 pathway into a different genomic background (Olofsson *et al.*, 2016). Indeed, some genes encoding enzymes of the C_4 pathway produce a sister relationship between C_3+C_4 individuals belonging to nuclear clade II and C_4 individuals (Dunning *et al.*, 2017), and the hybridization might have strengthened a weak C_4 pathway, highlighting the importance of hybridization for photosynthetic diversification (Kadereit *et al.*, 2017). After this introgression, the C_4 and C_3+C_4 lineages evolved independently despite their close geographic proximity. However, the C_4 accessions from Africa

possess alleles that group with C_3+C_4 individuals, while such alleles are at very low frequency in geographically distant C_4 accessions from Asia and Australia (Fig. 4b). This pattern suggests recurrent gene flow between C_3+C_4 and C_4 individuals, but also between C_3+C_4 and C_3 individuals from the northern end of the C_3 range, and between the different C_4 groups (Fig. 4). While these exchanges are frequent among diploids, more drastic genomic mixtures are observed within polyploids (Fig. 4).

Besides the polyploids previously reported in South Africa (Ellis, 1981; Liebenberg & Fossey, 2001; Lundgren *et al.*, 2015), we report here hexaploids from Zambia and Mozambique, and a dodecaploid from Cameroon (Table S1). Some of these polyploids present evidence of admixture between different groups (Fig. 4), a pattern previously reported for herbarium samples, although their ploidy level was unknown (Olofsson *et al.*, 2016). The genomic compositions differ among the polyploids, which are placed in different parts of the organelle phylogenies. In addition, some of the polyploids have unmatched plastid and mitochondrial genomes (Fig. S1). Specifically, one dodecaploid from Cameroon and hexaploids from Zambia, together with samples of unknown ploidy, present plastid sequences related to lineage E, but mitochondrial sequences related to lineage G. This incongruence probably results from organelle capture during allopolyploidization (Rieseberg & Soltis, 1991; Obbard *et al.*, 2006), and the cross-comparison of organelle and nuclear patterns suggests a minimum of three independent polyploidization events. First, the hexaploids from Mozambique and South Africa might have arisen from an admixture between C_4 nuclear clades III and IV. Alternatively, the admixture might have occurred before polyploidization, as it is also observed in C_4 diploids from Madagascar and to some extent Burkina Faso (Fig. 4). Second, the Cameroonian dodecaploid presents similar contributions of the three nuclear clades II, III and IV. Third, the Zambian hexaploids are mainly assigned to nuclear clade III, with contributions from nuclear clades II and IV. Based on the organelle phylogenies, polyploids might have arisen multiple times in the region, although secondary gene flow between different ploidy levels (e.g. via tetraploids) might explain the pattern. The Cameroonian dodecaploids and Zambian hexaploids might similarly share tetraploid or hexaploid ancestors. Recurrent allopolyploidization between lineages with distinct physiological backgrounds might have created further opportunities for niche expansion (e.g. Meimberg *et al.*, 2009), which can be illustrated here by the secondary migration of C_4 hexaploids to colder regions in South Africa.

Concluding remarks

Using phylogenomics of organelle and nuclear genomes, we obtain a detailed picture of the phylogeographic history of the grass *Alloteropsis semialata*, which is the only known species to encompass C_3 , C_3+C_4 and C_4 populations. The history of *A. semialata* is characterized by recurrent

divergence resulting from low seed dispersal followed by episodic pollen-mediated admixture. We suggest that these dynamics are responsible for the high genetic diversity observed within the species, which in turn facilitated the emergence of its unparalleled photosynthetic diversity. We detect multiple episodes of admixture between photosynthetically distinct genotypes, in multiple cases during segmental allopolyploidy. These results show that the photosynthetic types are not genetically isolated, but form a continuous species complex. Some of the secondary exchanges brought C_4 genes into C_3+C_4 populations, and we suggest that the mixing of C_4 components during admixture boosted physiological specialization. Importantly, our study reveals an unprecedented instance of unidirectional gene flow from a C_4 to a non- C_4 genome. In addition to boosting physiological innovation, secondary exchanges between previously isolated lineages therefore allowed the rapid spread of the novel C_4 trait into seeds inhabiting distant regions.

Acknowledgements

This study was funded by the European Research Council (grant ERC-2014-STG-638333), the Royal Society (grant RGF\EA\181050) and has benefited from “Investissements d’Avenir” grants managed by the Agence Nationale de la Recherche (CEBA, ref. ANR-10-LABX-25-01 and TULIP, ref. ANR-10-LABX-41). Edinburgh Genomics, which contributed to the sequencing, is partly supported through core grants from the NERC (R8/H10/56), MRC (MR/K001744/1), and BBSRC (BB/J004243/1). P.A.C. is funded by a Royal Society University Research Fellowship (grant URF\R\180022). We thank Pauline Raimondeau for lab support.

Author contributions

MEB, LTD, JKO, CPO and PAC designed the study. MEB did the phylogenetic analyses. LTD did the allele analyses. EVC did the population genomics analyses. JKO, SM and GB generated data. OH, RP, SM and IL generated the genome sizes. MRL did isotope analyses. MSV contributed with samples. MEB and PAC wrote the manuscript with the help of all co-authors.

References

- Aagesen L, Biganzoli F, Bena J, Godoy-Bürki AC, Reinheimer R, Zuloaga FO. 2016.** Macro-climatic distribution limits show both niche expansion and niche specialization among C_4 Panicoids. *PLoS ONE* **11**: e0151075.
- Andrews S. 2010.** FastQC: a quality control tool for high throughput sequence data. Available from: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

- Atkinson RRL, Mockford EJ, Bennett C, Christin P-A, Spriggs EL, Freckleton RP, Thompson K, Rees M, Osborne CP. 2016.** C₄ photosynthesis boosts growth by altering 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 B: Biological Sciences* **334**: 309–345.
- 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.
- Buggs RJA, Pannell JR. 2006.** Rapid displacement of a monoecious plant lineage is due to pollen swamping by a dioecious relative. *Current Biology* **16**: 996–1000.
- Burgess N, Hales JD, Underwood E, Dinerstein E, Olson D, Itoua I, Schipper J, Ricketts T, Newman K. 2004.** *Terrestrial ecoregions of Africa and Madagascar: A conservation assessment*. Washington, DC: Island Press.
- Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009.** trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972–1973.
- Christin P-A, Besnard G, Samaritani E, Duvall MR, Hodkinson TR, Savolainen V, Salamin N. 2008.** Oligocene CO₂ decline promoted C₄ photosynthesis in grasses. *Current Biology* **18**: 37–43.
- Christin P-A, Osborne CP, Sage RF, Arakaki M, Edwards EJ. 2011.** C₄ eudicots are not younger than C₄ monocots. *Journal of Experimental Botany* **62**: 3171–3181.
- Clark J, Hidalgo O, Pellicer J, Liu H, Marquardt J, Robert Y, Christenhusz M, Zhang S, Gibby M, Leitch IJ, et al. 2016.** Genome evolution of ferns: Evidence for relative stasis of genome size across the fern phylogeny. *New Phytologist* **210**: 1072–1082.
- Doležal 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.
- Doležal J, Greilhuber J, Suda J. 2007.** Estimation of nuclear DNA content in plants using flow cytometry. *Nature Protocols* **2**: 2233–2244.
- Droissart V, Dauby G, Hardy OJ, Deblauwe V, Harris DJ, Janssens S, Mackinder BA, Blach-Oergaard A, Sonké B, Sosef MSM, et al. 2018.** Beyond trees: Biogeographical regionalization of tropical Africa. *Journal of Biogeography* **35**: 1153–1167.
- Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. 2006.** Relaxed phylogenetics and dating with confidence. *PLoS Biology* **4**: 699–710.
- Drummond AJ, Rambaut A. 2007.** BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evolutionary Biology* **7**: 214.

539 **Dunning LT, Lundgren MR, Moreno-Villena JJ, Namaganda M, Edwards EJ, Nosil P,**
540 **Osborne CP, Christin P-A. 2017.** Introgression and repeated co-option facilitated the
541 recurrent emergence of C₄ photosynthesis among close relatives. *Evolution* **71**: 1541–1555.

542 **Dunning LT, Moreno-Villena JJ, Lundgren MR, Dionora J, Salazar P, Adams C, Nyirenda F,**
543 **Olofsson JK, Mapaura A, Grundy IM, et al. 2019a.** Key changes in gene expression
544 identified for different stages of C₄ evolution in *Alloteropsis semialata*. *Journal of*
545 *Experimental Botany* **70**: 3255–3268.

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

550 **Dupont LM, Behling H, Kim JH. 2008.** Thirty thousand years of vegetation development and
551 climate change in Angola (Ocean Drilling Program Site 1078). *Climate of the Past* **4**: 107–
552 124.

553 **Ebihara A, Ishikawa H, Matsumoto S, Lin SJ, Iwatsuki K, Takamiya M, Watano Y, Ito M.**
554 **2005.** Nuclear DNA, chloroplast DNA, and ploidy analysis clarified biological complexity of
555 the *Vandenboschia radicans* complex (Hymenophyllaceae) in Japan and adjacent areas.
556 *American Journal of Botany* **92**: 1535–1547.

557 **Edwards EJ, Still CJ. 2008.** Climate, phylogeny and the ecological distribution of C₄ grasses.
558 *Ecology Letters* **11**: 266–276.

559 **Edwards EJ, Osborne CP, Stromberg CAE, Smith SA, Bond WJ, Christin PA, Cousins AB,**
560 **Duvall MR, Fox DL, Freckleton RP, et al. 2010.** The origins of C₄ grasslands: Integrating
561 evolutionary and ecosystem science. *Science* **328**: 587–591.

562 **Ehleringer JR, Monson RK. 1993.** Evolutionary and ecological aspects of photosynthetic pathway
563 variation. *Annual Review of Ecology and Systematics* **24**: 411–439.

564 **Ellis RP. 1974.** The significance of the occurrence of both Kranz and non-Kranz leaf anatomy in
565 the grass species *Alloteropsis semialata*. *Journal of Science* **70**: 169–173.

566 **Ellis RP. 1981.** *Relevance of comparative leaf anatomy in taxonomic and functional research on*
567 *the South African Poaceae*. DSc thesis, University of Pretoria, South Africa.

568 **Emms DM, Kelly S. 2019.** OrthoFinder: Phylogenetic orthology inference for comparative
569 genomics. *Genome Biology* **20**: 238.

570 **Evanno G, Regnaut S, Goudet J. 2005.** Detecting the number of clusters of individuals using the
571 software STRUCTURE: A simulation study. *Molecular Ecology* **14**: 2611–2620.

572 **Folk RA, Mandel JR, Reudenstein JVF. 2017.** Ancestral gene flow and parallel organellar
573 genome capture result in extreme phylogenomic discord in a lineage of angiosperms.
574 *Systematic Biology* 66: 320–337.

575 **Frost P. 1996.** The ecology of miombo woodlands. In: Campbell B, ed. *The miombo in transition:*
576 *woodlands and welfare in Africa*. Bogor, Indonesia: Centre for International Forestry
577 Research, 11–57.

578 **Gibbs Russell GE. 1983.** The taxonomic position of C₃ and C₄ *Alloteropsis semialata* (Poaceae) in
579 southern Africa. *Bothalia* 14: 205–213.

580 **Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten**
581 **U, Putnam N, et al. 2012.** Phytozome: A comparative platform for green plant genomics.
582 *Nucleic Acids Research* 40: D1178–D1186.

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

586 **Hatch MD. 1987.** C₄ photosynthesis: a unique blend of modified biochemistry, anatomy and
587 ultrastructure. *Biophysica et Biochimica Acta - Reviews on Bioenergetics* 895: 81–106.

588 **Ivory SJ, Russell J. 2016.** Climate, herbivory, and fire controls on tropical African forest for the
589 last 60 ka. *Quaternary Science Reviews* 148: 101–114.

590 **Kadereit G, Bohley K, Lauterbach M, Tefarikis DT, Kadereit JW. 2017.** C₃-C₄ intermediates
591 may be of hybrid origin – a reminder. *New Phytologist* 215: 70–76.

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

594 **Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A,**
595 **Markowitz S, Duran C, et al. 2012.** Geneious Basic: An integrated and extendable desktop
596 software platform for the organization and analysis of sequence data. *Bioinformatics* 28:
597 1647–1649.

598 **Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I. 2015.** CLUMPAK: A
599 program for identifying clustering modes and packaging population structure inferences
600 across K. *Molecular Ecology Resources* 15: 1179–1191.

601 **Korneliussen TS, Albrechtsen A, Nielsen R. 2014.** ANGSD: Analysis of Next Generation
602 Sequencing Data. *BMC Bioinformatics* 15: 356.

603 **Langmead B, Salzberg SL. 2012.** Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9:
604 357–359.

605 **Larsson A. 2014.** AliView: A fast and lightweight alignment viewer and editor for large datasets.
606 *Bioinformatics* 30: 3276–3278.

- Lawton RM. 1978.** A study of the dynamic ecology of Zambian vegetation. *The Journal of Ecology* **66**: 175–198.
- Lee-Yaw JA, Grassa CJ, Joly S, Andrew RL, Rieseberg LH. 2019.** An evaluation of alternative explanations for widespread cytonuclear discordance in annual sunflowers (*Helianthus*). *New Phytologist* **221**: 515–526.
- Lehmann CER, Griffith DM, Simpson KJ, Anderson TM, Archibald S, Beerling DJ, Bond WJ, Denton E, Edwards EJ, Forrestel EJ, et al. 2019.** Functional diversification enabled grassy biomes to fill global climate space *bioRxiv* [Preprint] doi: 10.1101/583625.
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009.** The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.
- Liebenberg EJJ, Fossey A. 2001.** Comparative cytogenetic investigation of the two subspecies of the grass *Alloteropsis semialata* (Poaceae). *Botanical Journal of the Linnean Society* **137**: 243–248.
- Linder HP, de Klerk HM, Born J, Burgess ND, Fjeldsø J, Rahbek C. 2012.** The partitioning of Africa: Statistically defined biogeographical regions in sub-Saharan Africa. *Journal of Biogeography* **39**: 1189–1205.
- Long SP. 1983.** C₄ photosynthesis at low temperatures. *Plant, Cell & Environment* **6**: 345–363.
- Lorenzen ED, Heller R, Siegmund HR. 2012.** Comparative phylogeography of African savannah ungulates. *Molecular Ecology* **21**: 3656–3670.
- Loureiro J, Rodriguez E, Doležal 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, Lehmann CER, Chatelet DS, Kynast RG, Namaganda M, Vorontsova MS, Hall RC, Elia J, et al. 2015.** Photosynthetic innovation broadens the niche within a single species. *Ecology Letters* **18**: 1021–1029.
- Lundgren MR, Christin P-A, Escobar EG, Ripley BS, Besnard G, Long CM, Hattersley PW, Ellis RP, Leegood RC, Osborne CP. 2016.** Evolutionary implications of C₃–C₄ intermediates in the grass *Alloteropsis semialata*. *Plant, Cell & Environment* **39**: 1874–1885.
- Lundgren MR, Dunning LT, Olofsson JK, Moreno-Villena JJ, Bouvier JW, Sage TL, Khoshravesh R, Sultmanis S, Stata M, Ripley BS, et al. 2019.** C₄ anatomy can evolve via a single developmental change. *Ecology Letters* **22**: 302–312.
- Mairal M, Sanmartín I, Herrero A, Pokorny L, Vargas P, Aldasoro JJ, Alarcón M. 2017.** Geographic barriers and Pleistocene climate change shaped patterns of genetic variation in the Eastern Afrotropical biodiversity hotspot. *Scientific Reports* **7**, 45749.

- 641 **McDonough MM, Šumbera R, Mazoch V, Ferguson AW, Phillips CD, Bryja J. 2015.**
642 Multilocus phylogeography of a widespread savanna-woodland-adapted rodent reveals the
643 influence of Pleistocene geomorphology and climate change in Africa's Zambezi region.
644 *Molecular Ecology* **24**: 5248–5266.
- 645 **Meimberg H, Rice KJ, Milan NF, Njoku CC, McKay JK. 2009.** Multiple origins promote the
646 ecological amplitude of allopolyploid *Aegilops* (Poaceae). *American Journal of Botany* **96**:
647 1262–1273.
- 648 **Meisner J, Albrechtsen A. 2018.** Inferring population structure and admixture proportions in low-
649 depth NGS data. *Genetics* **210**: 719–731.
- 650 **Miller MA, Pfeiffer W, Schwartz T. 2010.** *Creating the CIPRES Science Gateway for inference of*
651 *large phylogenetic trees*. In: Proceedings of the Gateway Computing Environments
652 Workshop (GCE), New Orleans, LA, pp 1–8.
- 653 **Muñoz-Rodríguez P, Carruthers T, Wood JRI, Williams BRM, Weitemier K, Kronmiller B,**
654 **Ellis D, Anglin NL, Longway L, Harris SA, et al. 2018.** Reconciling conflicting
655 phylogenies in the origin of sweet potato and dispersal to Polynesia. *Current Biology* **28**:
656 1246–1256.
- 657 **Obbard DJ, Harris SA, Buggs RJA, Pannell JR. 2006.** Hybridization, polyploidy, and the
658 evolution of sexual systems in *Mercurialis* (Euphorbiaceae). *Evolution* **60**: 1801–1815.
- 659 **Obermayer R, Leitch IJ, Hanson L, Bennett MD. 2002.** Nuclear DNA C-values in 30 species
660 double the familial representation in pteridophytes. *Annals of Botany* **90**: 209–217.
- 661 **Olofsson JK, Bianconi M, Besnard G, Dunning LT, Lundgren MR, Holota H, Vorontsova**
662 **MS, Hidalgo O, Leitch IJ, Nosil P, et al. 2016.** Genome biogeography reveals the
663 intraspecific spread of adaptive mutations for a complex trait. *Molecular Ecology* **25**: 6107–
664 6123.
- 665 **Olofsson JK, Cantera I, Van de Paer C, Hong-Wa C, Zedane L, Dunning LT, Alberti A,**
666 **Christin P-A, Besnard G. 2019a.** Phylogenomics using low-depth whole genome
667 sequencing: A case study with the olive tribe. *Molecular Ecology Resources* **19**: 877–892.
- 668 **Olofsson JK, Dunning LT, Lundgren MR, Barton HJ, Thompson J, Cuff N, Ariyaratne M,**
669 **Yakandawala D, Sotelo G, Zeng K, et al. 2019b.** Population-specific selection on standing
670 variation generated by lateral gene transfers in a grass. *Current Biology* **29**: 3921–3927.
- 671 **Osborne CP, Salomaa A, Kluyver TA, Visser V, Kellogg EA, Morrone O, Vorontsova MS,**
672 **Clayton WD, Simpson DA. 2014.** A global database of C₄ photosynthesis in grasses. *New*
673 *Phytologist* **204**: 441–446.
- 674 **Patel RK, Jain M. 2012.** NGS QC toolkit: A toolkit for quality control of next generation
675 sequencing data. *PLoS ONE* **7**: e30619.

- 676 **Pearcy RW, Ehleringer J. 1984.** Comparative ecophysiology of C₃ and C₄ plants. *Plant, Cell &*
677 *Environment* **7**: 1–13.
- 678 **Pedersen CET, Albrechtsen A, Etter PD, Johnson EA, Orlando L, Chikhi L, Siegmund HR,**
679 **Heller R. 2018.** A southern African origin and cryptic structure in the highly mobile plains
680 zebra. *Nature Ecology and Evolution* **2**: 491–498.
- 681 **Quinlan AR, Hall IM. 2010.** BEDTools: A flexible suite of utilities for comparing genomic
682 features. *Bioinformatics* **26**: 841–842.
- 683 **R Core Team. 2018.** R: A language and environment for statistical computing. R Foundation for
684 Statistical Computing, Vienna, Austria. Available from: <https://www.R-project.org/>
- 685 **Rambaut A, Suchard MA, Xie W, Drummond AJ. 2013.** Tracer v1.6. Available from:
686 <http://tree.bio.ed.ac.uk/software/tracer/>
- 687 **Rieseberg LH, Soltis DE. 1991.** Phylogenetic consequences of cytoplasmic gene flow in plants.
688 *Evolutionary Trends in Plants* **5**: 65–84.
- 689 **Sage RF. 2004.** The evolution of C₄ photosynthesis. *New Phytologist* **161**: 341–370.
- 690 **Sage RF, Christin P-A, Edwards EJ. 2011.** The C₄ plant lineages of planet Earth. *Journal of*
691 *Experimental Botany* **62**: 3155–3169.
- 692 **Sage RF, Stata M. 2015.** Photosynthetic diversity meets biodiversity: The C₄ plant example.
693 *Journal of Plant Physiology* **172**: 104–119.
- 694 **Sayyari E, Mirarab S. 2016.** Fast coalescent-based computation of local branch support from
695 quartet frequencies. *Molecular Biology and Evolution* **33**: 1654–1668.
- 696 **Scheiter S, Higgins SI, Osborne CP, Bradshaw C, Lunt D, Ripley BS, Taylor LL, Beerling DJ.**
697 **2012.** Fire and fire-adapted vegetation promoted C₄ expansion in the late Miocene. *New*
698 *Phytologist* **195**: 653–666.
- 699 **Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. 2015.** BUSCO:
700 Assessing genome assembly and annotation completeness with single-copy orthologs.
701 *Bioinformatics* **31**: 3210–3212.
- 702 **Skotte L, Korneliussen TS, Albrechtsen A. 2013.** Estimating individual admixture proportions
703 from next generation sequencing data. *Genetics* **195**: 693–702.
- 704 **Smith BN, Brown WV. 1973.** The Kranz syndrome in the Gramineae as indicated by carbon
705 isotopic ratios. *American Journal of Botany* **60**: 505–513.
- 706 **Stamatakis A. 2014.** RAxML version 8: A tool for phylogenetic analysis and post-analysis of large
707 phylogenies. *Bioinformatics* **30**: 1312–1313.
- 708 **Toews DPL, Brelsford A. 2012.** The biogeography of mitochondrial and nuclear discordance in
709 animals. *Molecular Ecology* **21**: 3907–3930.

- Vincens A. 1991.** Late quaternary vegetation history of the South-Tanganyika basin. Climatic implications in South Central Africa. *Palaeogeography, Palaeoclimatology, Palaeoecology* **86**: 207–226.
- Vincens A, Tiercelin JJ, Buchet G. 2006.** New Oligocene-early Miocene microflora from the southwestern Turkana Basin. Palaeoenvironmental implications in the northern Kenya Rift. *Palaeogeography, Palaeoclimatology, Palaeoecology* **239**: 470–486.
- Watcharamongkol T, Christin P-A, Osborne CP. 2018.** C₄ photosynthesis evolved in warm climates but promoted migration to cooler ones. *Ecology Letters* **21**: 376–383.
- White F. 1983.** *The Vegetation of Africa*. Paris, France: UNESCO.
- Yemane K, Robert C, Bonnefille R. 1987.** Pollen and clay mineral assemblages of a late Miocene lacustrine sequence from the northwestern Ethiopian highlands. *Palaeogeography, Palaeoclimatology, Palaeoecology* **60**: 123–141.
- Zhang C, Rabiee M, Sayyari E, Mirarab S. 2018.** ASTRAL-III: Polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics* **19**: 153.

Figures

Fig. 1. Habitat diversity of *Alloteropsis semialata*. a) inflorescence of a C₄ hexaploid population in Mozambique (MOZ1601); b) tussocks of a C₃ diploid population in South Africa (RSA2); c) grassland in northern Australia where C₄ diploids can be found (population AUS4 in Olofsson *et al.*, 2019); d) C₄ hexaploids in the lowland dry miombo woodlands of northern Mozambique (MOZ1601); e) C₄ hexaploids in the highland wet miombo woodlands of northwestern Zambia (ZAM1507); f) miombo woodlands in central Tanzania, where diploid C₃+C₄ populations can be found (TAN2).

Fig. 2. Origin and dispersal of *Alloteropsis semialata* in Africa. The time-calibrated phylogenetic tree based on mitochondrial genomes is shown, with letters on nodes (A-G) indicating the organelle lineages (see Fig. S1 for branch support and 95% HPD intervals around estimated ages). All sampled African populations are shown on the map, with lineages coloured as in the mitochondrial phylogeny. Arrows indicate dispersal events.

Fig. 3. Nuclear history of *Alloteropsis*. The multigene coalescent species tree was estimated from 3,553 genome-wide nuclear markers. Pie charts on nodes indicate the proportion of quartet trees that support the main (dark grey), first (pale blue) and second (light grey) alternative topologies. Dashed-line pie charts indicate main topologies with local posterior probability < 0.95. Branch lengths are in coalescent units, except the terminal branches, which are arbitrary. Roman numbers I–IV denote the four main nuclear clades of *A. semialata*, which are indicated with coloured shades. Major geographic regions are indicated.

Fig. 4. Genetic structure and genomic composition of *Alloteropsis semialata*. a) Individual-based admixture analysis assigning individuals to the two best numbers of clusters (K) (see Fig. S4). Nuclear clades (I-IV) and organelle lineages (A-G) are indicated at the bottom. b) Genome composition obtained from phylogenetic analyses using alleles of single-copy exons. Colours in bars represent the proportion of genomic reads assigned to each of the four major nuclear clades of *A. semialata* (see Table S2).

Fig. 5. Putative history of the C_4 nuclear genome of *Alloteropsis semialata*. The C_4 nuclear genome is shown in orange, on top of the seed history. Gene flow between lineages is indicated by horizontal connections.

Supporting Information

Fig. S1. Time-calibrated phylogeny of *Alloteropsis semialata* based on plastid (left) and mitochondrial (right) genome sequences. Brown bars on nodes indicate 95% HPD. Circles highlight nodes with posterior probability ≥ 0.95 . The organelle clades are indicated with letters (A-G), and the coloured shades correspond to nuclear clades (Fig. 3).

Fig. S2. Multigene coalescent species trees estimated from two datasets with less stringent filtering; a) alignments not trimmed (4,345 genes retained); b) sites with more than 70% of missing data were discarded (4,198 genes retained). Pie charts on nodes indicate the proportion of quartet trees that support the main (dark grey), first (pale blue) and second (light grey) alternative topologies. Local posterior probabilities are indicated near nodes. Branch lengths are in coalescent units.

Fig. S3. Multigene coalescent species tree based on 3,553 genes. The gene set is as in Fig. 3, but only diploid individuals were retained. Pie charts on nodes indicate the proportion of quartet trees that support the main (dark grey), first (pale blue) and second (light grey) alternative topologies. Local posterior probabilities are indicated near nodes. Branch lengths are in coalescent units.

Fig. S4. Genetic structure of *Alloteropsis semialata*. a) Principal component analysis using genome-wide nuclear data. Colours correspond to major nuclear clades as identified in Fig. 3; b) Ranked eigenvectors of principal component analysis; c) (top) Mean likelihood ($\pm$ SD) over five runs for each number of clusters tested for the admixture analysis (K :1-10); (bottom) fit improvement for the admixture analysis, as calculated according to Evanno *et al.* (2005).

Table S1. Sample information (.xlsx).

Table S2. Genome composition analysis.

Fig. 1

bioRxiv preprint doi: <https://doi.org/10.1101/2020.04.23.053280>; this version posted April 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fig. 4

bioRxiv preprint doi: <https://doi.org/10.1101/2020.04.23.053280>; this version posted April 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Fig. 5

bioRxiv preprint doi: <https://doi.org/10.1101/2020.04.23.053280>; this version posted April 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.