OXFORD

G3
Genes | Genomes | Genetics

A haplotype-resolved chromosome-level genome assembly of *Urochloa decumbens* cv. Basilisk resolves its allopolyploid ancestry and composition

Camilla Ryan,[1] Fiona Fraser,[1] Naomi Irish,[1] Tom Barker,[1] Vanda Knitlhoffer,[1] Alex Durrant,[1] Gillian Reynolds (iD),[1]
Gemy Kaithakottil,[1] David Swarbreck (iD),[1] Jose J. De Vega (iD)[1,*]

[1]Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, UK

*Corresponding author: Earlham Institute, Norwich Research Park, Norwich NR4 7UZ, UK. Email: jose.devega@earlham.ac.uk

Haplotyped-resolved phased assemblies aim to capture the full allelic diversity in heterozygous and polyploid species to enable accurate genetic analyses. However, building non-collapsed references still presents a challenge. Here, we used long-range interaction Hi-C reads (high-throughput chromatin conformation capture) and HiFi PacBio reads to assemble the genome of the apomictic cultivar Basilisks from *Urochloa decumbens* ($2n = 4x = 36$), an outcrossed tetraploid Paniceae grass widely cropped to feed livestock in the tropics. We identified and removed Hi-C reads between homologous unitigs to facilitate their scaffolding and employed methods for the manual curation of rearrangements and misassemblies. Our final phased assembly included the 4 haplotypes in 36 chromosomes. We found that 18 chromosomes originated from diploid *Urochloa brizantha* and the other 18 from either *Urochloa ruziziensis* or diploid *U. decumbens*. We also identified a chromosomal translocation between chromosomes 5 and 32, as well as evidence of pairing exclusively within subgenomes, except for a homoeologous exchange in chromosome 21. Our results demonstrate that haplotype-aware assemblies accurately capture the allelic diversity in heterozygous species, making them the preferred option over collapsed-haplotype assemblies.

Keywords: *Urochloa decumbens*; Hi-C; HiFi; genome assembly; polyploid

Introduction

Livestock contributes to the livelihoods of more than two-thirds of the world's rural poor (FAO 2024). The scarcity and seasonal availability of forage and its low nutritional value are the main limiting constraints in meat and milk production in tropical regions (Bonilla-Cedrez *et al.* 2023). The use of improved forage cultivars can contribute to building resilience in pasture-based food systems, while boosting animal welfare and the income of rural families. In the broader context, livestock has a large land and carbon footprint, and more nutritious grass can result in lower methane and nitrous oxide emissions per weight (Ferreira *et al.* 2021; Bonilla-Cedrez *et al.* 2023). Intensification can also reduce unnecessary degradation of natural land and ecosystem services (Jank *et al.* 2014).

Native to Africa, *Urochloa* species were introduced into South America in the 1970s because of their good carrying capacity, nutritional value, grazing tolerance, and adaptability to areas of low fertility (Miles *et al.* 2004; Ferreira *et al.* 2021). Their broad usage in South America and their ability to interspecifically hybridize have enabled the development of high-performing *Urochloa* cultivars using recurrent selection breeding programs at CIAT and EMBRAPA (Pizarro *et al.* 2013; Maass *et al.* 2015). Improved *Urochloa* varieties have been estimated to double the number of livestock units per area and year compared to natural pastures (Miles *et al.* 2004; Jank *et al.* 2014).

Further advances in producing new cultivars are hindered by the genera's complex genomic architecture, such as variable ploidy, complex phylogenies, apomixis, and a lack of genomic resources (Higgins *et al.* 2022; Tomaszewska *et al.* 2023). There are no assemblies for any polyploid *Urochloa* species despite their agronomic importance and the benefits this could bring to breeding programs (Ferreira *et al.* 2021). Only single-haplotype genome assemblies were available for the diploid species *U. ruziziensis* ($2n = 18$) (Worthington *et al.* 2021), which has limited agronomic interest, prior to this study. These collapsed assemblies are likely not an accurate representation of the species' haplotypic diversity due to the heterozygous outcrossed nature of the genus.

Until recently, most genome assemblies were collapsed into single haplotypes because algorithms could not distinguish between allelic haplotypes when assembling short-read sequences (Whibley *et al.* 2021). However, recent advances in sequencing technologies, namely highly accurate long reads and long-range interaction information from chromosomal conformation capture (Hi-C) methods, are reducing the complexity of genome assembly and enabling the production of haplotype-aware assemblies, including heterozygous and polyploid species (Li and Durbin 2024).

Haplotype-aware, chromosome-level assemblies can greatly benefit crop breeding programs by enabling more accurate population structure and marker-trait association studies and a better understanding of gene dosage effects and allelic-driven complex traits, such as apomixis and self-incompatibility (Njaci *et al.* 2023).

This study presents the first haplotype-aware chromosome-level assembly of the polyploid *U. decumbens* ($2n = 4x = 36$),

particularly from a widely used apomictic cultivar named Basilisk. It demonstrates the feasibility and value of producing fully haplotype-resolved assemblies in heterozygous tetraploid species. We also extended our genome analysis by identifying structural features of *U. decumbens* and clarifying the species' much-discussed ancestry (Higgins *et al.* 2022; Masters *et al.* 2024; Tomaszewska *et al.* 2023).

Methods

Sample collection

U. decumbens cv. Basilisk plants were grown for 8 weeks before DNA extraction from seeds originating from Uganda accessed via the Australian Pastures Genebank (APG 58378) (also known as CIAT 606). After 8 weeks, leaf material was harvested following 2 days in dark conditions. The same leaf tissue and genotype were used for all the work described below.

DNA extraction

High molecular weight DNA extraction was performed using the Nucleon PhytoPure kit, with a slightly modified version of the recommended protocol. One gram of leaf material was ground under liquid nitrogen for a total grinding time of 9–10 min. Following this, the powder was thoroughly resuspended (more aggressively than indicated by the manufacturer protocol) using a 10-mm bacterial spreader loop. This method of homogenate mixing was used for all subsequent mixing steps before the addition of the chloroform and resin. After the ice incubation, 300 μL of resin was added along with the chloroform. Three hundred microliters is at the upper end of the recommended range. The chloroform extraction was followed by extraction with 25:24:1 phenol:chloroform:isoamyl alcohol, which was added to the previous upper phase, mixed at 4°C on a 3D platform rocker for 10 min and then centrifuged at 3,000*g* for 10 min. The upper phase from this procedure was then transferred to a 15-mL Falcon tube and precipitated as recommended by the manufacturer's protocol. The final elution was left open in a fume hood for 2 h to allow residual phenol and ethanol to evaporate, and the DNA sample was left at room temperature overnight.

Generating HiFi reads

The library for this project was constructed at the Earlham Institute, Norwich, UK, using the SMRTbell Express Template Prep Kit 2.0 (PacBio, P/N 100-983-900). 12.6 μg of sample was manually sheared with the Megaruptor 3 instrument (Diagenode, P/N B06010003). The sample underwent AMPure PB bead (PacBio, P/N 100-265-900) purification and concentration before undergoing library preparation using the SMRTbell Express Template Prep Kit 2.0 (PacBio, P/N 100-983-900). The HiFi library was prepared according to the HiFi protocol version 03 (PacBio, P/N 101-853-100), and the final library was size fractionated using the SageELF system (Sage Science, P/N ELF0001), 0.75% cassette (Sage Science, P/N ELD7510). The library was quantified by fluorescence (Invitrogen Qubit 3.0, P/N Q33216), and the size of fractions was estimated from a smear analysis performed on the FEMTO Pulse System (Agilent, P/N M5330AA). The loading calculations for sequencing were completed using the PacBio SMRT Link Binding Calculator v10.2. Sequencing primer v5 was annealed to the adapter sequence of the HiFi library. The library was bound to the sequencing polymerase with the Sequel II Binding Kit v2.2 (PacBio, P/N 102-089-000). Calculations for primer and polymerase binding ratios were kept at default values for the library type. Sequel II DNA internal control 1.0 was spiked into the library at the standard concentration prior to sequencing. The sequencing chemistry used was Sequel II Sequencing Plate 2.0 (PacBio, P/N 101-820-200) and the Instrument Control Software v10.1.0.125432. The library was sequenced on 3 Sequel II SMRT Cell 8M. The parameters for sequencing per SMRT cell were as follows: adaptive loading default settings, 30-h movie, 2-h pre-extension time, and 80 pM on plate loading concentration.

Generating Hi-C reads

Sample material for the Omni-C library prep was 100 mg of *U. decumbens* young leaf tissue that was harvested, snap-frozen in liquid nitrogen, and stored at −80°C. The Omni-C library was prepared using the Dovetail Omni-C Kit (SKU: 21005) according to the manufacturer's protocol for "Non-mammal v1.2B". Briefly, the chromatin was fixed with disuccinimidyl glutarate and formaldehyde in the nucleus. The cross-linked chromatin was then digested in situ with DNase I (0.05 μL). Following digestion, the cells were lysed with SDS to extract the chromatin fragments, which were bound to chromatin capture beads. Next, the chromatin ends were repaired and ligated to a biotinylated bridge adapter followed by proximity ligation of adapter-containing ends. After proximity ligation, the cross-links were reversed, the associated proteins were degraded, and the DNA was purified then converted into a sequencing library [NEBNext Ultra II DNA library Prep Kit for Illumina (E7645)] using Illumina-compatible adaptors [NEBNext Multiplex Oligos for Illumina (Index Primers Set 1) (E7335)]. Biotin-containing fragments were isolated using streptavidin beads prior to PCR amplification.

The library pool was diluted to 0.5 nM using EB (10 mM Tris pH8.0) in a volume of 18 μL before spiking in 1% Illumina phiX Control v3. This was denatured by adding 4-μL 0.2N NaOH and incubating at room temperature for 8 min, after which it was neutralized by adding 5-μL 400 mM Tris pH 8.0. A master mix of EPX1, EPX2, and EPX3 from Illumina's Xp 2-lane kit v1.5 (20043130, Illumina) was made and 63 μL added to the denatured pool leaving 90 μL at a concentration of 100 pM. This was loaded onto a NovaSeq SP flow cell using the NovaSeq Xp Flow Cell Dock. The flow cell was then loaded onto the NovaSeq 6000 along with a NovaSeq 6000 SP cluster cartridge, buffer cartridge, and 300-cycle SBS cartridge (20028400, Illumina). The NovaSeq had NVCS v1.7.5 and RTA v3.4.4 and was set up to sequence 150-bp PE reads. The data were demultiplexed and converted to fastq format using Illumina Bcl2Fastq2.

Genome assembly

The workflow used to produce the genome assembly is represented in Fig. 1. Firstly, a unitig assembly was produced using HiFiasm v0.18 (Cheng *et al.* 2021, 2022). HiFiasm produces multiple assemblies with increasing contiguity by iteratively improving the assembly graph and increasingly discarding (or collapsing) minor variations. We decided to advance with the unitig assembly (instead of contigs) for scaffolding because unitigs are haplotype specific (Supplementary Table 1); therefore, the unitig assembly included all 4 haplotypes. The quality of the unitig assembly was comparable to that of the contig-level assembly, and the number of unitigs assembled was within the scaffolder's processing limits (Zhou *et al.* 2022).

Omni-C reads were mapped to the unitig assembly, removed not primary and supplementary alignment (SAM flags 2304), and the read alignment file was pruned for the use in scaffolding. Pruning was first suggested by Zhang *et al.* (2018, 2019) and referred to the method of removing uninformative interchromosomal links between allelic haplotypes, which otherwise may
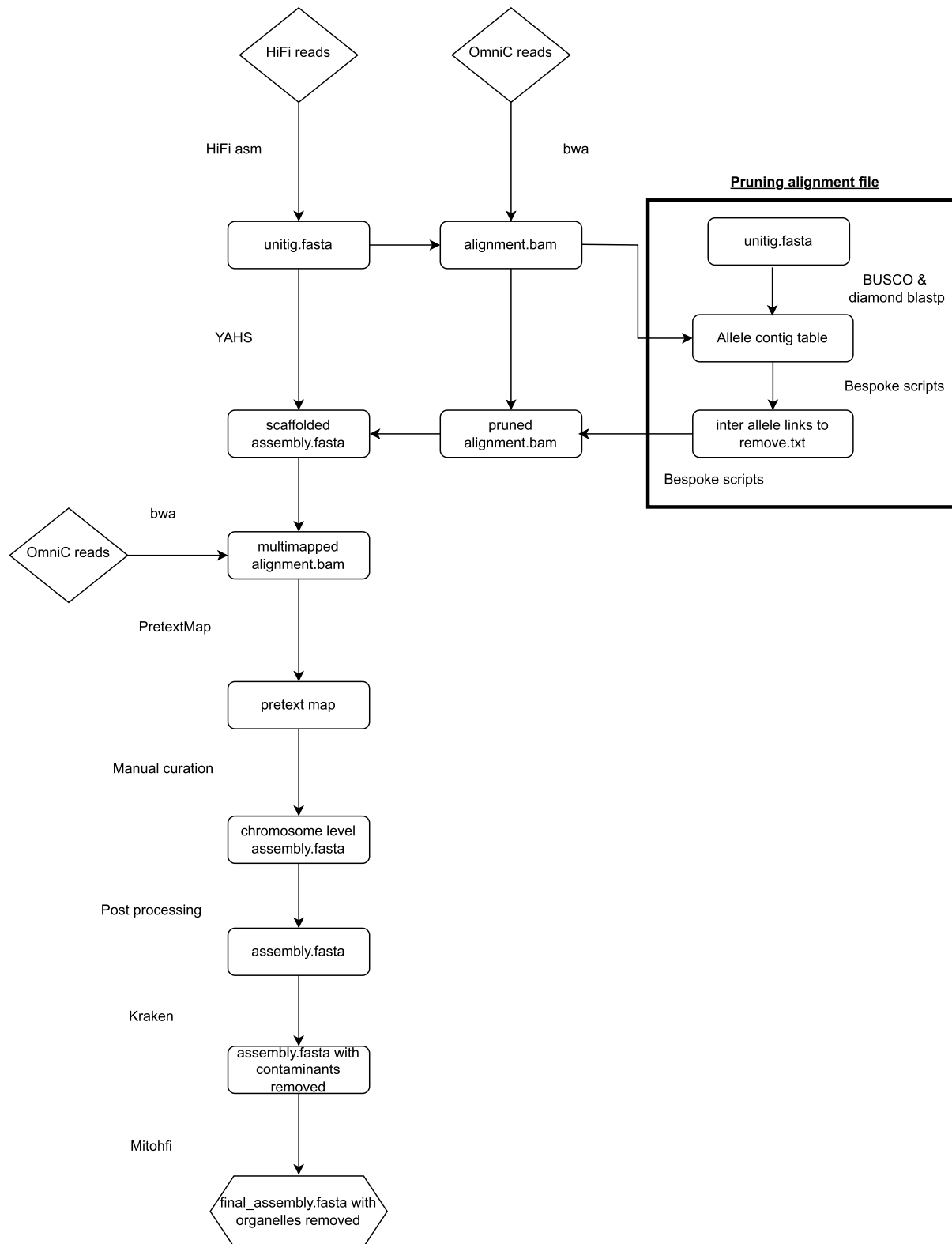
Fig. 1. Overview of the bioinformatics pipeline used to create the haplotype-resolved chromosome-level assembly of the allotetraploid *U. decumbens*.

result in misjoins during scaffolding. For that, the single-copy markers for the grasses family from BUSCO (Manni *et al.* 2021) were mapped with Diamond blastp (Buchfink *et al.* 2021) to the unitigs to generate a table of unitigs with the same single-copy marker (named allelic unitigs). Then, the Omni-C reads linking allelic unitigs were removed. The pruned Omni-C alignment file was used with YAHS v1.2a.2 (Zhou *et al.* 2022) to scaffold the unitig assembly with a minimum mapping quality of 30 and the following lists of resolutions: 1,000, 2,000, 5,000, 10,000, 20,000, 50,000, 100,000, 200,000, 500,000, 1,000,000, 2,000,000, 5,000,000,

10,000,000, 20,000,000, 50,000,000, 100,000,000, 200,000,000, and 500,000,000.

The scaffolded assembly was manually curated following a workflow used by the Darwin Tree of Life program developed by the Sanger Institute (Cambridge, UK). Firstly, the Omni-C reads were remapped to the scaffolded assembly allowing for multimapping reads; although multimapping reads are not informative during scaffolding, they are useful during manual curation, e.g. to identify regions with repeat sequences. PretextMap v0.1.9 (Tree of Life Programme 2024a) was used to create a contact matrix (or pretext map) compatible with PretextView v0.2.5 (Tree of Life Programme 2024b), which is an interactive viewer that allows the editing of the contact matrix directly. Tracks for telomere sequences, gaps, and coverage were added to the pretext map to help with manual curation. The location of telomere sequences in the assembly was identified using tidk v0.2.31 (Brown *et al.* 2023); as the telomere sequences for *U. decumbens* are unknown, the motif "TTTAGGG" was used (Peska and Garcia 2020). Seqtk cutN v1.2 (Li, 2024) was used to localize gaps in the scaffolded genome over 10 bp. Finally, HiFi reads were mapped back to the scaffolded genome using Minimap2 v2.24 (Li 2018, 2021) to generate the coverage tracks. The alignment file was sorted with SAMtools v1.10 (Danecek *et al.* 2021) and used to calculate coverage across the genome using Mosdepth v0.3.3 (Pedersen and Quinlan 2018). Tracks were then created from bedgraphs and added to the pretext map using PretextGraph v0.0.6 (Tree of Life Programme 2024c). This coverage plot was overlayed onto the PretextGraph and allowed us to identify any large abnormalities in coverage—such as duplications that could then be manually removed.

After manual curation, the script rapid_pretext2tpf_XL.py (Tracey and Wood 2024) was used to create a new tpf file and regenerate the genome's Fasta file. Chromosome-length sequences were sorted and renamed based on homology to each other and subgenome ancestry and scaffolds by length.

Contaminant and organelle sequence removal

Kraken 2 v2.0.7 (Wood *et al.* 2019) and the "Standard-16 nucleotide database" version 2.0.7_refseq-201910 (Langmead 2024) were used to identify any scaffolds that did not belong to Viridiplantae. MitoHiFi v3.0.0 (Uliano-Silva *et al.* 2023) was used to identify scaffolds belonging to mitochondria (mtdna) and chloroplast (pltd) and the most similar sequences available to assemble the 2 organelles. For *U. decumbens*, the most similar mtdna found was from *Microstegium vimineum* (NC_072666.1; accessed May 24, 2023) and the most similar pltd came from *U. decumbens* (NC_030066.1; accessed May 24, 2023). Scaffolds identified as non-Viridiplantae (contaminants), chloroplast, or mitochondrial were removed from the final assembly.

Genome quality assessment

The quality of the final assembly was assessed using several measures; basic assembly metrics, including contiguity, were produced using Abyss v1.9.0 (Simpson *et al.* 2009), and assembly completeness was assessed from BUSCO v5.3.2 (Manni *et al.* 2021) analysis using the Poales v10 database. Merqury v1.3 (Rhie *et al.* 2020) was used to produce Kmer completeness metrics and Kmer spectra plots to ensure the assembly captured all the content from the reads and to produce consensus quality (QV) metrics. QV measures likely assembly errors based on Kmers found only in the reads. This is then converted into a Phred-equivalent score (Rhie *et al.* 2020). The final chromosomes were mapped, using Minimap2 v2.24 (Li 2018, 2021), to an existing reference of the closely related diploid *U. ruziziensis* (GCA_015476505.1; accessed March, 2023).

Gene annotation

Gene models were generated from the *U. decumbens* assembly using *Robust and Extendable eukaryotic Annotation Toolkit* (REAT) v0.6.1 (Earlham Institute 2024a) and Minos v1.8.0 (Earlham Institute 2024b), which are pipelines that used Mikado v2.3.4 (Venturini *et al.* 2018), Portcullis v1.2.4 (Mapleson *et al.* 2018), and multiple third-party tools (listed in the above repositories) as dependencies. Identification of repetitive elements was performed using the EI-Repeat pipeline v1.3.4 (Earlham Institute 2024c), which masked the genome assembly using RepeatMasker v4.0.7 (Tarailo-Graovac and Chen 2009) and the RepBase database and a de novo repeat database constructed with RepeatModeler v1.0.11 (Smit and Hubley 2008). REAT's "transcriptomic workflow" was used for alignment of short-read RNA-seq data generated in a previous study (Higgins *et al.* 2022) using HISAT2 v2.2.1 (Kim *et al.* 2019) with high-confidence splice junctions identified using Portcullis v.1.2.4 (Mapleson *et al.* 2018). Alignments from short reads were assembled using StringTie v2.1.5 (Kovaka *et al.* 2019) and Scallop v0.10.5 (Shao and Kingsford 2017). A consolidated set of transcriptome-derived gene models was generated using Mikado v2.3.3 (Venturini *et al.* 2018). REAT's "homology workflow" was used to align protein sequences from 7 related species (Supplementary Table 2) against the *U. decumbens* assembly. Proteins were aligned using Spaln v2.4.7 (Gotoh 2008) and filtered to remove misaligned proteins. The same proteins were also aligned using miniprot v0.3 (Li 2023) and filtered. The aligned proteins from both alignment methods were clustered into loci and a consolidated set of gene models derived with Mikado v2.3.4. REAT's "prediction workflow" was used to generate a set of evidence-guided gene predictions by training Augustus (Stanke and Morgenstern 2005) with high-confidence gene models from the previous workflows. Four alternative Augustus runs were performed with varying weightings of evidence, which were provided to EVidenceModeler (Haas *et al.* 2008) along with the transcriptome and protein evidence to generate consensus gene structures. Genes were also predicted using Helixer (Holst *et al.* 2023), a deep neural network approach, using its publicly available plant model. The final set of gene models was selected using Minos from the outputs from REAT's homology, transcriptome, and prediction workflows, plus Helixer's gene models. Gene models were classified as coding, noncoding, or transposable, and with a high- or low-confidence score, based on the support from RNA-seq or protein evidence (from the 7 related species plus UniProt's Magnoliopsida proteins) with previously defined criteria (Grewal *et al.* 2024).

Ancestry analysis

Sourmash v4.8.5 (Irber *et al.* 2024) was used to generate Kmer signatures for all *U. decumbens* and *U. ruziziensis* chromosomes, perform pairwise comparisons between genomes, and plot dendrograms and heatmaps. Kmer composition and frequency signatures, comparisons, and plots were generated for Kmer sizes 3–21 in increments of 2 and 21–161 in increments of 10. Subgenomic clustering was determined using the "strict cut" criteria in Reynolds *et al.* (2024). In short, subgenomic clusters are deemed correct if, and only if, all chromosomes belonging to a subgenome are within the same cluster. In the plots, all chromosome numbers are annotated with their subgenome ancestry. *U. ruziziensis* chromosomes (GCA_015476505.1; accessed March

2023) were also annotated with introgression information and number.

Reads from 3 *Urochloa* species were aligned to the final assembly to clarify its genome composition and ancestry. Reads from the diploid *U. decumbens* were downloaded from NCBI's sequence read archive (SRR16327313; accessed on May 22, 2023). Reads from the diploid *Urochloa brizantha* were kindly shared by EMBRAPA (M. Pessoa, *per. Comm.*). Reads from *U. ruziziensis* were downloaded from PRJNA437375. Each set of reads was mapped using Minimap2 v2.24 (Li 2018), and the coverage was plotted in R v3.6.0 (R Core Team 2021) using a modified version of the function plot_coverage() from PafR v0.0.2 (Winter 2020).

Assessing repeat content

Transposable elements were identified from the genome de novo using EDTA v2.1.0 (Ou *et al.* 2019). Long-terminal repeats (LTRs) were extracted from the output of EDTA, and the distribution and density of intact LTRs were plotted across the genome in R v3.6.0 (R Core Team 2021).

Table 1. Completeness and contiguity metrics for the final curated genome assembly of *U. decumbens*.

Statistics	Complete genome (n = 7,122)	Chromosomes only (n = 36)
Total assembly size (Gb)	2.879	2.474
N50 contig length (Mb)	66.6	69.83
Max contig length (Mb)	104.3	104.3
Complete BUSCOs	4,860 (99.2%)	4,859 (99.2%)
Complete and single BUSCOs	36 (0.7%)	59 (1.2%)
Complete and duplicated BUSCOs	4,824 (98.5%)	4,800 (98.0%)
Fragmented BUSCOs	5 (0.1%)	5 (0.1%)
Missing BUSCOs	31 (0.7%)	32 (0.7%)
Merqury QV	67.326	71.832
Merqury completeness	97.859	95.689

Identifying structural changes through synteny

Structural changes were initially identified using the coverage plots and later in greater detail using a syntenic approach. For that, high-confidence protein-coding genes were selected from the annotation results. However, only those coding genes found on chromosomes (i.e. not on scaffolds) were retained. A table of all-vs-all protein alignments was created using DIAMOND blastp v2.0.15 (Buchfink *et al.* 2021). This table of homologous genes and the genome annotation was used with MCScanx v2 (Wang 2022) to identify putative syntenic chromosomic regions and produce a collinearity file. The results of MCScanX were plotted using SynVisio (Bandi and Gutwin 2020).

Results and discussion

Capturing the full allelic diversity in a heterozygous and polyploid grass species

A haplotype-resolved chromosome-level de novo assembly from *U. decumbens* cultivar Basilisk was generated using a combination of HiFi reads and Omni-C data. Starting from 10,806 unitig sequences generated by the assembler (N50 3.6 Mb, 3.03 Gb total; Supplementary Table 1), 85.9% of the assembly (2.55 Gb, 3,727 unitigs) was later successfully anchored into 36 chromosomes and 7,086 unlocalized scaffolds (Table 1; Supplementary Table 3). The 36 chromosomes contained 99.2% complete BUSCO markers (Supplementary Table 1). A Kmer spectra of the HiFi reads vs the assembly evidence the assembly accurately reflected the raw read content (Supplementary Fig. 1).

Chromosomes were numbered according to the contact matrix after manual curation, which allowed us to identify pairs of chromosomes organized into subgenomes (Fig. 2). The distinctive "chain" pattern of contacts between homologous chromosomes within a subgenome and a faint signal between homoeologous pairs indicated an allotetraploid composition, i.e. preferential pairing restricted within subgenomes and no evidence of homoeologous exchanges in the contact matrix (Fig. 2).
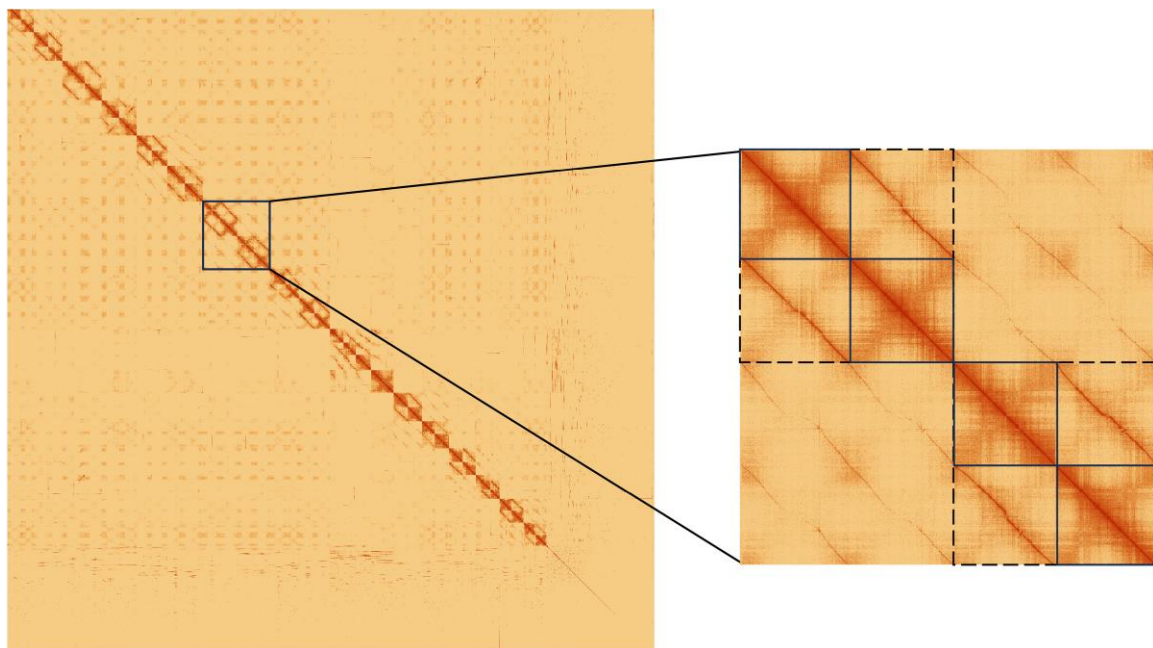

Fig. 2. Hi-C contact matrix following manual curation. A close-up of chromosomes 13–16 shows the visual differences between subgenomes (dashed lines) and individual chromosomes (solid lines) within and among subgenomes.
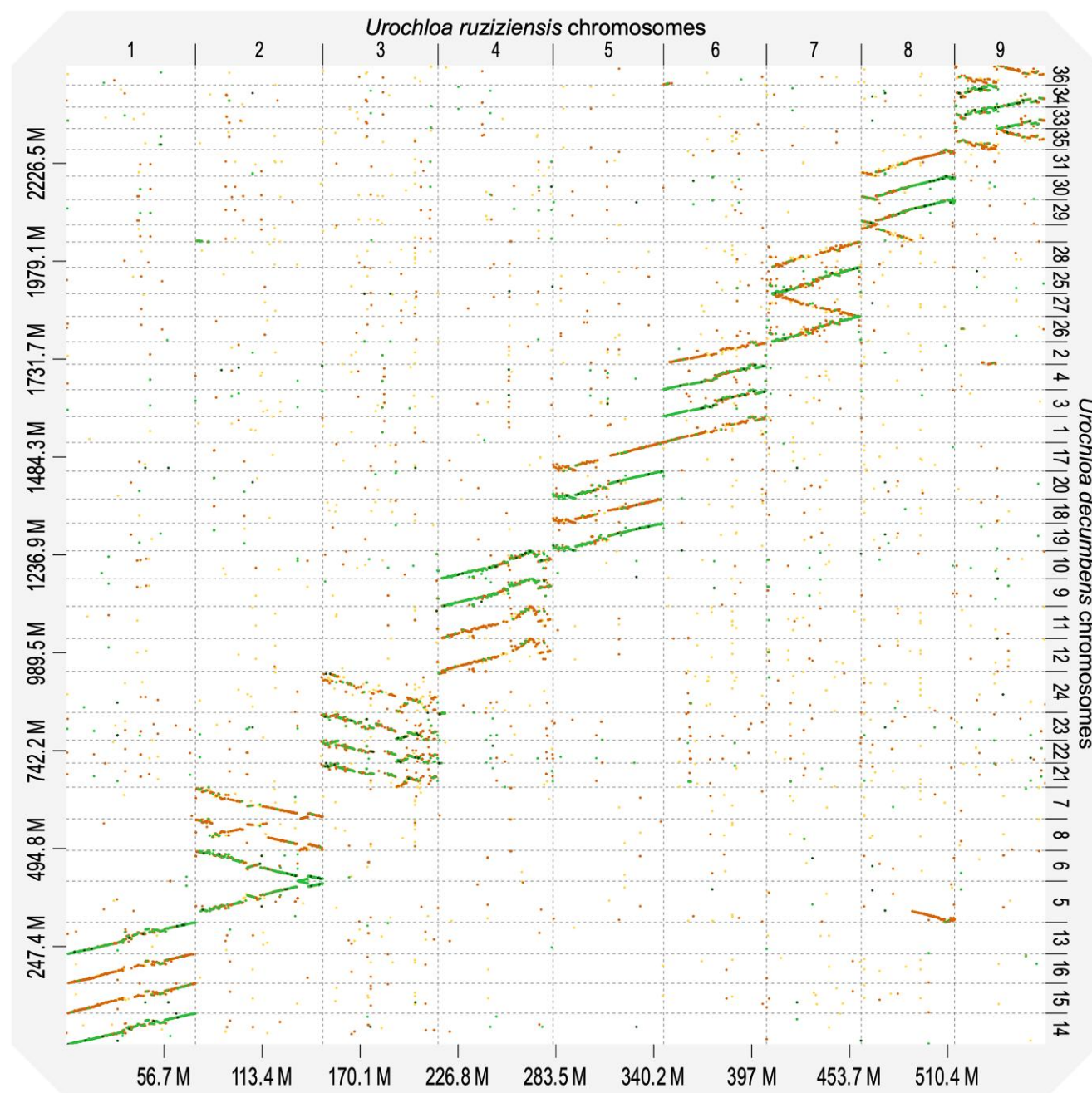
Fig. 3. Dotplot representation of the alignments from the 36 chromosomes from allotetraploid *U. decumbens* (4n = 4x = 36) to the 9 chromosomes from the single-haplotype assembly of diploid *U. ruziziensis* (2n = 18). Sequence colors represent alignment identity: 0.00–0.25, yellow; 0.25–0.50, orange; 0.50–0.75, light green; and 0.75–1.00, dark green.

When we aligned the genome to the single-haploid assembly of *U. ruziziensis*, we observed all 4 *U. decumbens* haplotypes aligned to each *U. ruziziensis* haplotype (Fig. 3). For each chromosome, 2 of the 4 haplotypes had a lower identity to *U. ruziziensis* than the other 2, indicating these chromosomes likely derived from *U. ruziziensis* or a closely related ancestor to it. We also observed evidence of translocations in the dotplot.

Out of a total of 4,896 BUSCO markers for Poales, 99.2% were found complete, and 98.5% were duplicated. The frequency of alignments showed that most markers aligned 4 times—which is expected in a complete tetraploid assembly (Fig. 4). Similar values were obtained only considering markers found within

chromosomes. This indicated that most genic content was captured in the anchored chromosomes. The REAT annotation pipeline predicted 126,000 protein-coding genes and 167,192 transcripts (accounting for alternative splicing) with a mean coding sequence (CDS) length of 1.7 kb. The final proteome (126,000 proteins) was also assessed with BUSCO and found to be of high quality [C: 99.9% (S: 0.1%, D: 99.8%), F: 0.0%, M: 0.1%, n: 4,896].

Ancestry of tetraploid *U. decumbens*

U. decumbens' chromosomes clustered in 2 groups by subgenome ancestry (Fig. 5a) based on Kmer frequency for most of
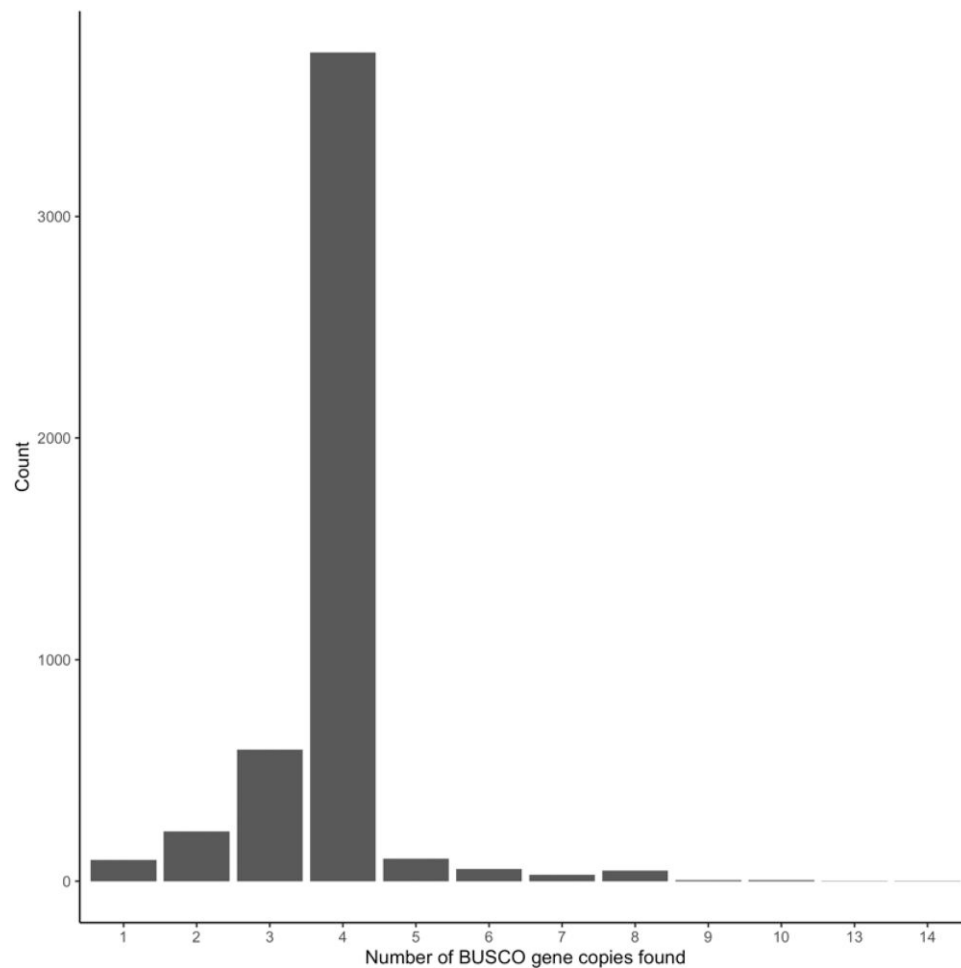
Fig. 4. Histogram of the number of times BUSCO single-copy markers aligned in the assembly (chromosomes only). Most single-copy markers were found 4 times, once per haplotype, as expected in a heterozygous tetraploid.
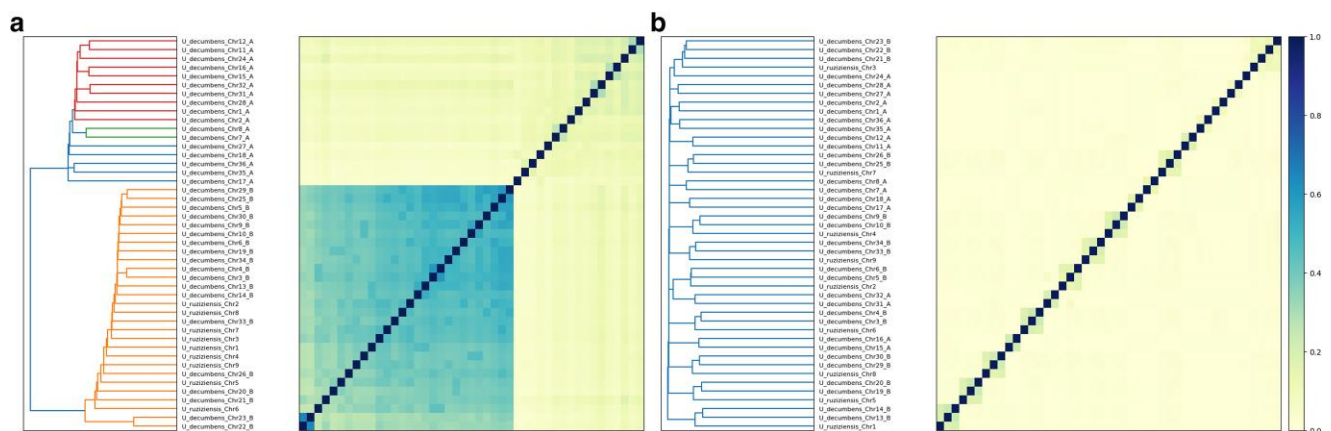

Fig. 5. a) Kmer frequency ($K = 81$) clustered *U. decumbens* by subgenome (labeled as a and b) and placed all *U. ruziziensis* chromosomes with one of the subgenomes b) However, chromosomes did not cluster by subgenome when using Kmer composition ($K = 81$) instead they evidence preferential pairing restricted within subgenomes, expect in chromosome 21 due to a chromosome exchange (also observed by coverage analysis). Branch and heatmap color represent sequence similarity based on Kmer signatures.

the sampled Kmer sizes ($K = 17$, 21–161; Supplementary Table 4). The genomes did not cluster by subgenomes based on Kmer composition (Fig. 5b) of any Kmer size. Instead, *U. decumbens*' dendrograms reflected the prevalence of preferential pairing within subgenome, except chromosome 21 (Fig. 5b); i.e. chromosomes clustered in pairs between homologous chromosomes (e.g. Chr13_B and Chr14_B) within the same subgenome.
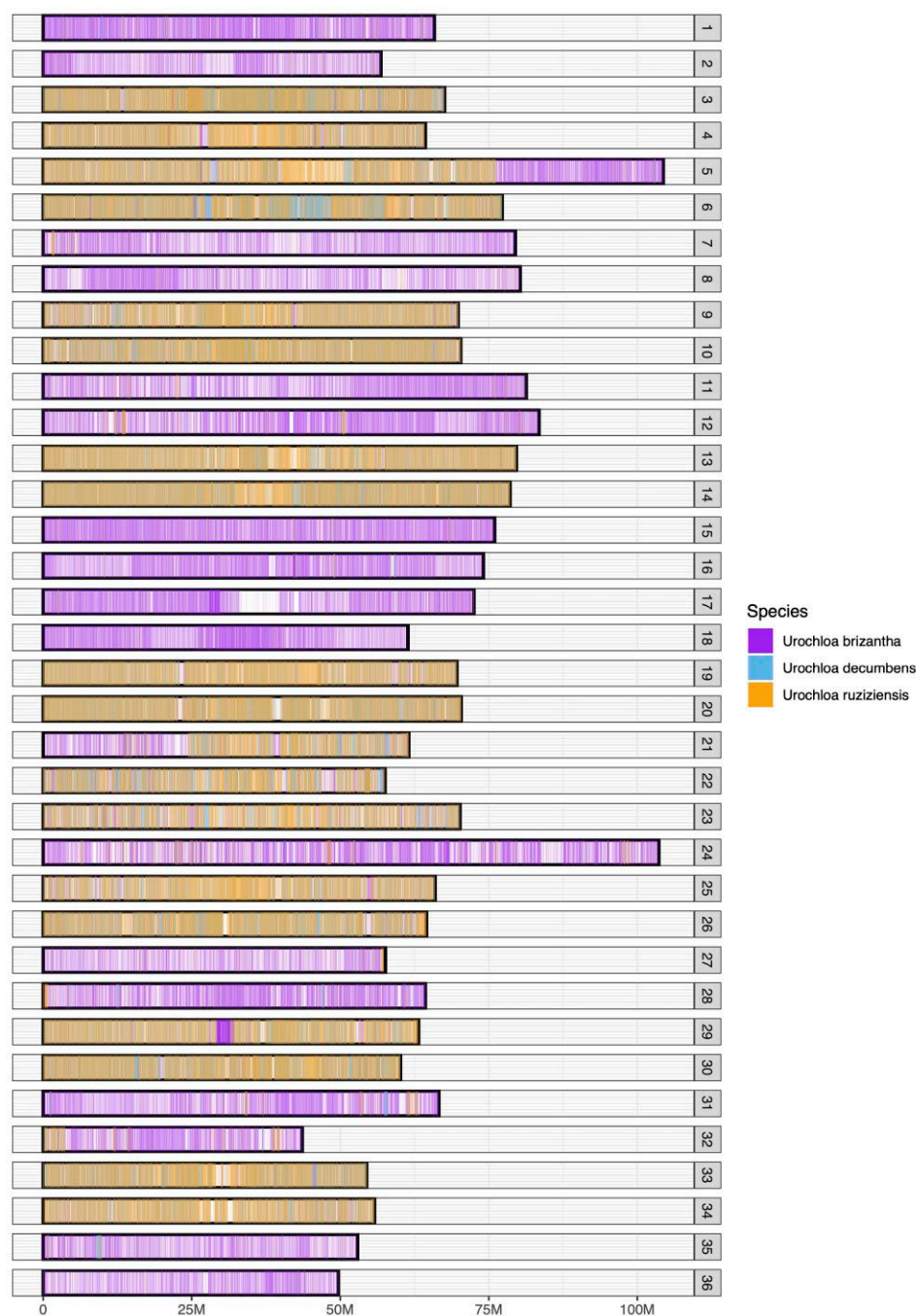
Fig. 6. Coverage(read depth) following the alignment from 3 diploid *Urochloa* species to the *U. decumbens* genome. Half of the chromosomes were contributed from *U. brizantha* (purple) or a close ancestor and the other half from diploid *U. decumbens* (blue) or *U. ruziziensis* (orange) or a close ancestor. Evidence for homoeologous exchange (chromosome 21) and translocations (chromosomes 5 and 32) can also be observed.

When we grouped the chromosomes from the *U. decumbens* and *U. ruziziensis* genomes together (Fig. 5), we noticed that all the *U. ruziziensis* chromosomes clustered with half of the chromosomes that also displayed higher similarity to *U. ruziziensis* in the dotplot. This provides further evidence supporting the relationship of one of the subgenomes in the allotetraploid *U. decumbens* with the diploid *U. ruziziensis*, while the other subgenomes have a different origin.

To infer the ancestry of each chromosome, we independently aligned whole-genome short reads (WGS) from diploid *U. ruziziensis*, diploid *U. decumbens*, and diploid *U. brizantha* to the new

assembly: *U. ruziziensis* and *U. decumbens* aligned to the same 18 chromosomes (Fig. 6; Supplementary Table 5), while *U. brizantha* reads aligned to the other 18 chromosomes (Fig. 6). We concluded that half the chromosomes' ancestry was from *U. brizantha*, while the other half was from either *U. ruziziensis*, diploid *U. decumbens*, or their common ancestor. We could not distinguish between these 2 species, as there was no difference between where reads from diploid *U. decumbens* and *U. ruziziensis* aligned (Supplementary Fig. 2).

Previous phylogenetic and ancestry analyses have shown diploid *U. decumbens* more closely related to diploid *U. ruziziensis*
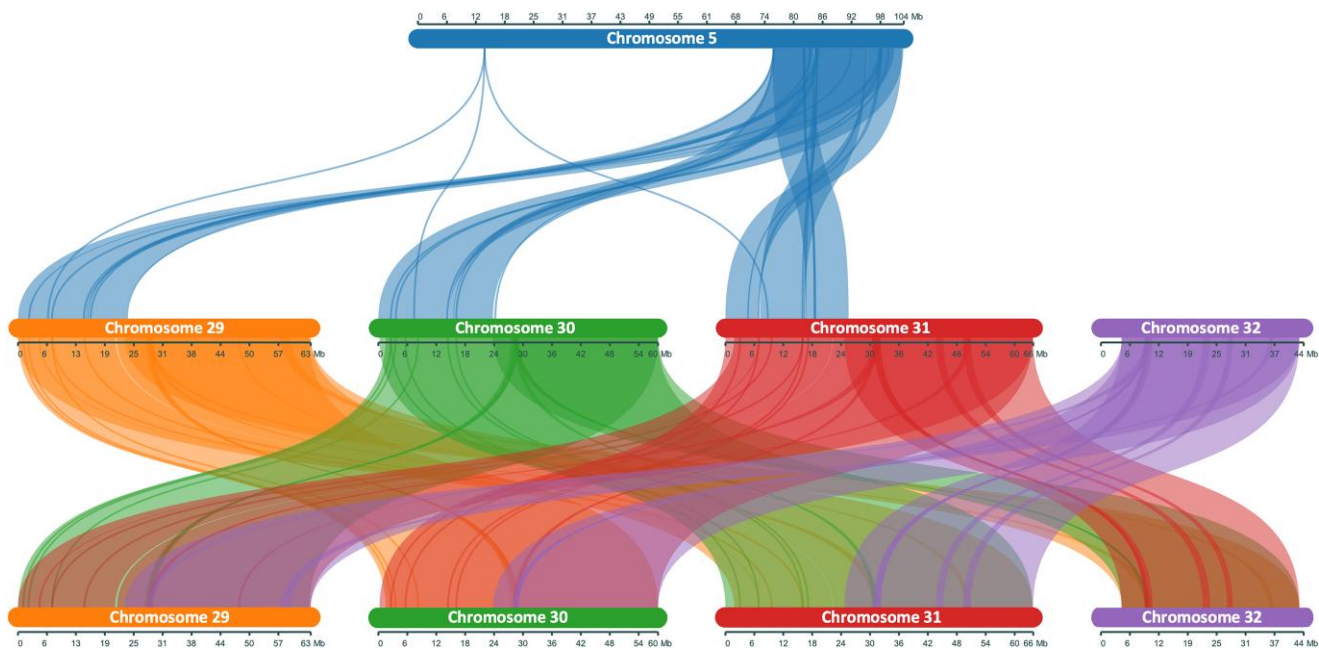
Fig. 7. Synteny between chromosomes evidences a translocation between the 5-end of chromosome 32 and the 3-end of chromosome 5. Synteny is conserved between chromosomes 29, 30, and 31, but not 32.

Table 2. Summary of repeat families regions found in the genome assembly.

Class	Count	Basepairs	Percentage
Long tandem repeats (LTRs)			
Copia	250,646	247,799,132	8.61
Gypsy	461,898	677,745,937	23.54
Unknown	405,127	424,732,747	14.75
Terminal inverted repeats (TIRs)			
CACTA	210,015	102,760,281	3.57
Mutator	181,651	70,232,389	2.44
PIF_harbinger	106,529	36,236,564	1.26
Tc1_mariner	163,519	46,738,014	1.62
hAT	572,94	20,684,037	0.72
Non-TIR			
Helitron	712,792	341,135,300	11.85
Total interspersed repeats	2,549,471	1,968,064,401	68.35

than polyploid *U. decumbens* (Higgins *et al.* 2022). Another study using fluorescence in situ hybridization proposed an ancestry of 9 chromosomes from *U. brizantha*, 9 chromosomes from *U. decumbens*, and 18 chromosomes from *U. ruziziensis* (Tomaszewska *et al.* 2023). This result likely reflects the difficulty of designing markers that do not cross-hybridize among these highly related species in the *Urochloa* species complex.

On the other hand, it had been suggested that tetraploid *U. decumbens* was a segmental allopolyploid based on genetic mapping (Worthington *et al.* 2016). However, we did not observe evidence of frequent pairing across subgenomes and homoeologous exchanges in the contact matrix (Fig. 2) or coverage plot (Fig. 6) to justify its classification as segmental allotetraploid. The only evidence of exchange between homoeologous pairs is chromosome 21 (Supplementary Table 3), as observed in Kmer composition analysis (Fig. 5b) and coverage analysis (Fig. 6). Chromosome 21 corresponds to chromosome 8 in *Setaria italica*. This is the same base

chromosome (chromosome 8) detected in the genetic maps in Worthington *et al.* (2016) that we think led to *U. decumbens*' "historical" classification as segmental allotetraploid. However, our results support this is exclusive to cv. Basilisk and not the whole species.

Finally, we also observed 1 translocation between chromosomes 5 and 32, where the beginning of chromosome 32 (*U. brizantha* ancestry) had been translocated to the end of chromosome 5 (*U. decumbens/ruziziensis* ancestry) in the cultivar Basilisk, which was collected from the wild and consequently reflects the variation to be expected in a wild apomictic lineage in this complex (Fig. 7).

Finally, EDTA predicted 2,549,471 interspersed repeats covering 68.35% of the genome (Table 2). The distribution of LTRs across the chromosomes also supported the division of the assembly into its distinct subgenomes, with homologous chromosomes sharing a more similar pattern of repeats than homoeologous chromosomes (Fig. 8), except in chromosomes 21–24, where we previously identified a homoeologous exchange in chromosome 21.

Conclusion

In this study, we have produced a haplotype-aware chromosome-level assembly of the heterozygous allotetraploid *U. decumbens* cv. Basilisk, an apomictic genotype, using HiFi PacBio long reads and Hi-C reads. These technologies have enabled the assembly of all 36 chromosomes of *U. decumbens* at a contiguity functional to the agronomic and scientific community. We also validated the removal (pruning) of Hi-C links between allelic haplotypes, which were innovatively detected using single-copy BUSCO markers, facilitating the anchoring of this haplotype-aware polyploid genome. Furthermore, this haplotype-aware assembly allowed us to identify the ancestry of each subgenome within *U. decumbens*. We concluded that the allotetraploid *U. decumbens* resulted from the hybridization of diploids from *U. brizantha* and either *U. ruziziensis* or *U. decumbens*. Furthermore, we did not find supporting evidence for its classification as a segmental allopolyploid but
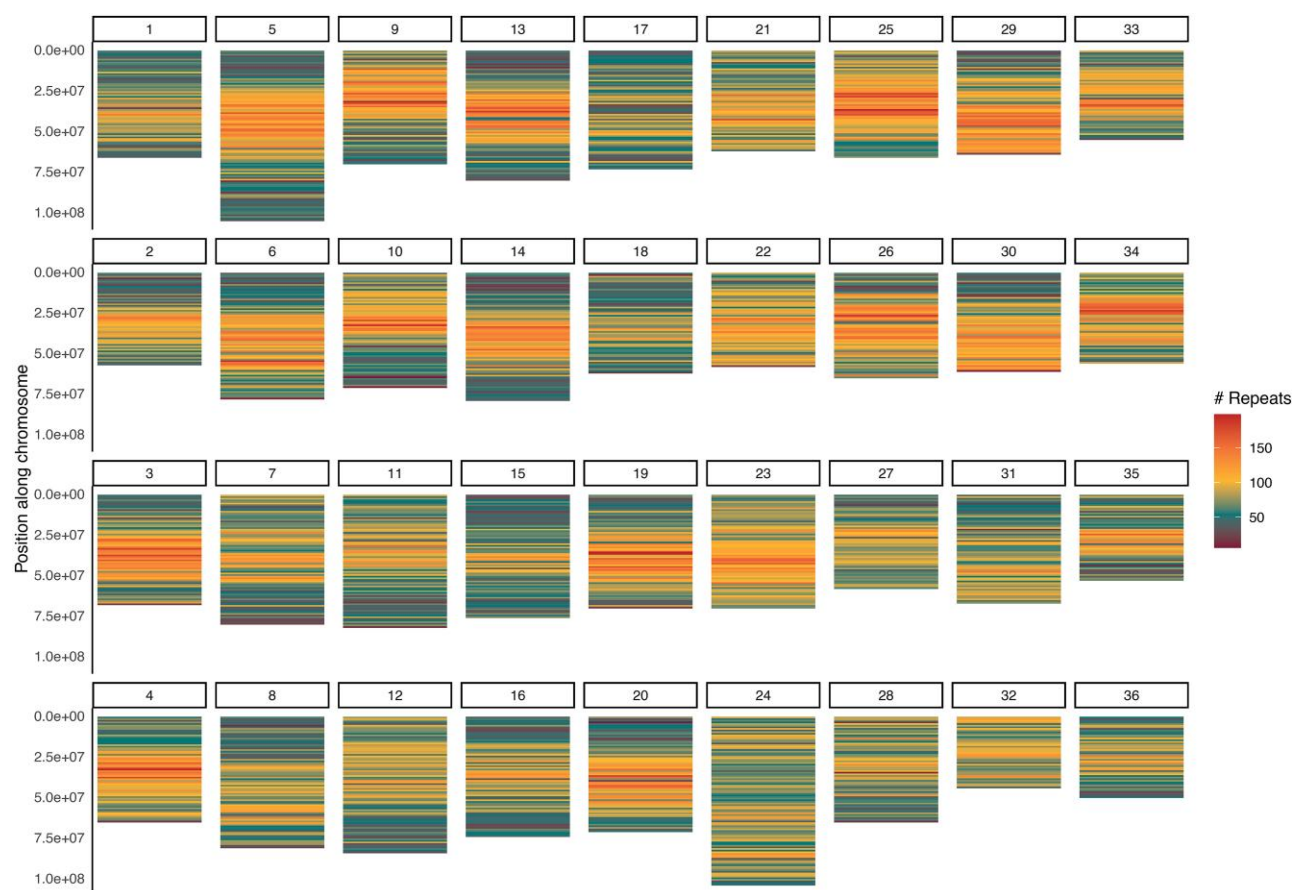
Fig. 8. Density of LTRs across the 36 assembled chromosomes evidenced repeat patterns were similar in chromosomes from the same subgenome, except in chromosomes 21–24 due to a homoeologous exchange.

for the nominal preferential pairing within subgenomes in allopolyploids. Finally, we believe only haplotype-aware assemblies accurately capture the allelic diversity in heterozygous species, and they should be the preferred option over collapsed-haplotype assemblies in the future.

Data availability

Raw reads are deposited in the SRA under accession PRJEB73762. The genome assembly, together with its gene annotation, was deposited in ENA with accession GCA_964030465.3 (https://www.ebi.ac.uk/ena/browser/view/GCA_964030465.3). The scripts used in this study are publicly available in GitHub (https://github.com/DeVegaGroup/HaplotypeAwareChromosomeLevelAssemblyUrochloaDecumbens).

Supplemental material available at G3 online.

Acknowledgments

All the authors contributed and approved this manuscript. The authors would like to acknowledge the support of the Norwich Bioscience Institutes Research Computing team. They would also like to thank the Technical Genomics group at the Earlham Institute, especially Drs. Sacha Lucchini, Kendall Baker, Leah Catchpole, Karim Gharbi, and Chris Watkins for their end-to-end support and roles in project administration and supervision. Additionally, the authors express their gratitude to Prof. Neil Hall, Dr. Christine Fosker, and their teams, for their roles in

Funding

Conflicts of interest

The authors declare no conflict of interest.

Literature cited

Bandi V, Gutwin C. 2020. Interactive exploration of genomic conservation. In: Graphics Interface 2020. https://openreview.net/forum?id=7-C5VJWbnI. [accessed 2024 May 24].

Bonilla-Cedrez C, Steward P, Rosenstock TS, Thornton P, Arango J, Kropff M, Ramirez-Villegas J. 2023. Priority areas for investment in more sustainable and climate-resilient livestock systems. Nat Sustain. 6(10):1279–1286. doi:10.1038/s41893-023-01161-1.

Buchfink B, Reuter K, Drost H-G. 2021. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 18(4):366–368. doi:10.1038/s41592-021-01101-x.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. 2021. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 18(2):170–175. doi:10.1038/s41592-020-01056-5.

Cheng H, Jarvis ED, Fedrigo O, Koepfli K-P, Urban L, Gemmell NJ, Li H. 2022. Haplotype-resolved assembly of diploid genomes without parental data. Nat Biotechnol. 40(9):1332–1335. doi:10.1038/s41587-022-01261-x.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, Whitwham A, Keane T, McCarthy SA, Davies RM, *et al.* 2021. Twelve years of SAMtools and BCFtools. GigaScience. 10(2):giab008. doi:10.1093/gigascience/giab008.

Earlham Institute. 2024a. EI-REPEAT. https://github.com/EICoreBioinformatics/eirepeat/. [accessed 2024 May 24].

Earlham Institute. 2024b. MINOS. https://github.com/EICoreBioinformatics/minos/. [accessed 2024 May 24].

Earlham Institute. 2024c. REAT. https://github.com/EI-CoreBioinformatics/reat/. [accessed 2024 May 24].

FAO. 2024. FAOSTAT. https://www.fao.org/faostat/en/#data. [accessed 2024 May 24].

Ferreira RCU, da Costa Lima Moraes A, Chiari L, Simeão RM, Vigna BBZ, de Souza AP. 2021. An overview of the genetics and genomics of the *Urochloa* species most commonly used in pastures. Front Plant Sci. 12:770461. doi:10.3389/fpls.2021.770461.

Gotoh O. 2008. A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. Nucleic Acids Res. 36(8):2630–2638. doi:10.1093/nar/gkn105.

Grewal S, Yang C, Scholefield D, Ashling S, Ghosh S, Swarbreck D, Collins J, Yao E, Sen TZ, Wilson M, *et al.* 2024. Chromosome-scale genome assembly of bread wheat's wild relative *Triticum timopheevii*. Sci Data. 11(1):420. doi:10.1038/s41597-024-03260-w.

Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. 9(1):R7. doi:10.1186/gb-2008-9-1-r7.

Higgins J, Tomaszewska P, Pellny TK, Castiblanco V, Arango J, Tohme J, Schwarzacher T, Mitchell RA, Heslop-Harrison JS, De Vega JJ, *et al.* 2022. Diverged subpopulations in tropical *Urochloa* (*Brachiaria*) forage species indicate a role for facultative apomixis and varying ploidy in their population structure and evolution. Ann Bot. 130(5):657–669. doi:10.1093/aob/mcac115.

Holst F, Bolger A, Günther C, Maß J, Triesch S, Kindel F, Kiel N, Saadat N, Ebenhöh O, Usadel B, *et al.* 2023. Helixer–de novo prediction of primary eukaryotic gene models combining deep learning and a hidden Markov mode. bioRxiv 2023.02.06.527280. https://doi.org/10.1101/2023.02.06.527280, preprint : not peer reviewed.

Brown M, González De la Rosa PM, Mark B. 2023. A telomere identification toolkit [dataset]. Zenodo. https://doi.org/10.5281/zenodo.10091385.

Irber L, Pierce-Ward NT, Abuelanin M, Alexander H, Anant A, Barve K, Baumler C, Botvinnik O, Brooks P, Dsouza D, *et al.* 2024.

Sourmash v4: a multitool to quickly search, compare, and analyze genomic and metagenomic data sets. J Open Source Softw. 9(98):6830. doi:10.21105/joss.06830.

Jank L, Barrios Sanzio C, Do Valle CB, Simeão RM, Alves GF. 2014. The value of improved pastures to Brazilian beef production. Crop Pasture Sci. 65(11):1132–1137. doi:10.1071/CP13319.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 37(8):907–915. doi:10.1038/s41587-019-0201-4.

Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 20(1):278. doi:10.1186/s13059-019-1910-1.

Langmead B. Kraken 2, KrakenUniq and Bracken indexes. https://benlangmead.github.io/aws-indexes/k2. [accessed 2024 May 24].

Li H. seqtk. https://github.com/lh3/seqtk. [accessed 2024 May 24].

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 34(18):3094–3100. doi:10.1093/bioinformatics/bty191.

Li H. 2021. New strategies to improve Minimap2 alignment accuracy. Bioinformatics. 37(23):4572–4574. doi:10.1093/bioinformatics/btab705.

Li H. 2023. Protein-to-genome alignment with miniprot. Bioinformatics. 39(1):btad014. doi:10.1093/bioinformatics/btad014.

Li H, Durbin R. 2024. Genome assembly in the telomere-to-telomere era. Nat Rev Genet. 25(9):658–670. doi:10.1038/s41576-024-00718-w.

Maass BL, Midega CAO, Mutimura M, Rahetlah VB, Salgado P, Kabirizi JM, Khan ZR, Ghimire SR, Rao IM. 2015. Homecoming of *Brachiaria*: improved hybrids prove useful for African animal agriculture. East Afr Agric For J. 81(1):71–78. doi:10.1080/00128325.2015.1041263.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. 2021. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol. 38(10):4647–4654. doi:10.1093/molbev/msab199.

Mapleson D, Venturini L, Kaithakottil G, Swarbreck D. 2018. Efficient and accurate detection of splice junctions from RNA-seq with Portcullis. GigaScience. 7(12):giy131. doi:10.1093/gigascience/giy131.

Masters LE, Tomaszewska P, Schwarzacher T, Zuntini AR, Heslop-Harrison P, Vorontsova MS. 2024. Phylogenomic analysis reveals five independently evolved African forage grass clades in the genus Urochloa. Ann Bot. 133(5–6):725–742. https://doi.org/10.1093/aob/mcae022

Miles JW, Do Valle CB, Rao IM, Euclides VPB. 2004. Brachiariagrasses. In: Moser LE, Burson BL, Sollenberger LE, editors. Warm-Season (C4) Grasses. John Wiley & Sons, Ltd. p. 745–783.

Njaci I, Waweru B, Kamal N, Muktar MS, Fisher D, Gundlach H, Muli C, Muthui L, Maranga M, Kiambi D, *et al.* 2023. Chromosome-level genome assembly and population genomic resource to accelerate orphan crop lablab breeding. Nat Commun. 14(1):1915. doi:10.1038/s41467-023-37489-7.

Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, Lugo CSB, Elliott TA, Ware D, Peterson T, *et al.* 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 20(1):275. doi:10.1186/s13059-019-1905-y.

Pedersen BS, Quinlan AR. 2018. Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics. 34(5):867–868. doi:10.1093/bioinformatics/btx699.

Peska V, Garcia S. 2020. Origin, diversity, and evolution of telomere sequences in plants. Front Plant Sci. 11:117. doi:10.3389/fpls.2020.00117.

Pizarro EA, Hare MD, Mutimura M, Bai C. 2013. Brachiaria hybrids: potential, forage use and seed yield. Trop Grassl-Forrajes Trop. 1(1):31–35. doi:10.17138/tgft(1)31-35.

R Core Team. 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/. [accessed 2024 May 24].

Reynolds G, Mumey B, Strnadova-Neeley V, Lachowiec J. 2024. Hijacking a rapid and scalable metagenomic method reveals sub-genome dynamics and evolution in polyploid plants. Appl Plant Sci. 12(4):e11581. doi:10.1002/aps3.11581.

Rhie A, Walenz BP, Koren S, Phillippy AM. 2020. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 21(1):245. doi:10.1186/s13059-020-02134-9.

Shao M, Kingsford C. 2017. Accurate assembly of transcripts through phase-preserving graph decomposition. Nat Biotechnol. 35(12): 1167–1169. doi:10.1038/nbt.4020.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol İ. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res. 19(6):1117–1123. doi:10.1101/gr.089532.108.

Smit AF, Hubley R. 2008. RepeatModeler Open-1.0. http://www.Repeatmasker.Org. [accessed 2024 May 24].

Stanke M, Morgenstern B. 2005. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 33(Web Server):W465–W467. doi:10.1093/nar/gki458.

Tarailo-Graovac M, Chen N. 2009. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. 25(1):4.10.1–4.10.14. doi:10.1002/0471250953.bi0410s25.

Tomaszewska P, Vorontsova MS, Renvoize SA, Ficinski SZ, Tohme J, Schwarzacher T, Castiblanco V, de Vega JJ, Mitchell RAC, Heslop-Harrison JSP. 2023. Complex polyploid and hybrid species in an apomictic and sexual tropical forage grass group: genomic composition and evolution in *Urochloa* (*Brachiaria*) species. Ann Bot. 131(1):87–108. doi:10.1093/aob/mcab147.

Tracey A, Wood J. 2024. [accessed 2024 May 24]. https://gitlab.com/wtsi-grit/rapid-curation/-/blob/0316318e208f9d5d9ac72c3a1da97e00d822e7ad/rapid_pretext2tpf_XL.py.

Tree of Life Programme. 2024a. PretextMap. https://github.com/sanger-tol/PretextMap. [acccesed 2024 May 24].

Tree of Life Programme. 2024b.PretextView. https://github.com/sanger-tol/PretextView. [accessed 2024 May 24].

Tree of Life Programme. 2024c. PretextGraph. https://github.com/sanger-tol/PretextGraph. [accessed 2024 May 24].

Uliano-Silva M, Ferreira JGRN, Krasheninnikova K, Blaxter M, Mieszkowska N, Hall N, Holland P, Durbin R, Richards T, Kersey P, *et al.* 2023. MitoHiFi: a Python pipeline for mitochondrial genome assembly from PacBio high fidelity reads. BMC Bioinformatics. 24(1):288. doi:10.1186/s12859-023-05385-y.

Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D. 2018. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. GigaScience. 7(8):giy093. doi:10.1093/gigascience/giy093.

Wang Y. 2022. MCScanX: Multiple Collinearity Scan toolkit X version. https://github.com/wyp1125/MCScanX. [accessed 2024 May 24].

Whibley A, Kelley JL, Narum SR. 2021. The changing face of genome assemblies: guidance on achieving high-quality reference genomes. Mol Ecol Resour. 21(3):641–652. doi:10.1111/1755-0998.13312.

Winter D. 2020. pafr. https://github.com/dwinter/pafr/. [accessed 2024 May 24].

Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. Genome Biol. 20(1):257. doi:10.1186/s13059-019-1891-0.

Worthington M, Heffelfinger C, Bernal D, Quintero C, Zapata YP, Perez JG, De Vega J, Miles J, Dellaporta S, Tohme J. 2016. A parthenogenesis gene candidate and evidence for segmental allopolyploidy in apomictic *Brachiaria decumbens*. Genetics. 203(3): 1117–1132. doi:10.1534/genetics.116.190314.

Worthington M, Perez JG, Mussurova S, Silva-Cordoba A, Castiblanco V, Cardoso Arango JA, Jones C, Fernandez-Fuentes N, Skot L, Dyer S, *et al.* 2021. A new genome allows the identification of genes associated with natural variation in aluminium tolerance in *Brachiaria* grasses. J Exp Bot. 72(2):302–319. doi:10.1093/jxb/eraa469.

Zhang J, Zhang X, Tang H, Zhang Q, Hua X, Ma X, Zhu F, Jones T, Zhu X, Bowers J, *et al.* 2018. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. Nat Genet. 50(11): 1565–1573. doi:10.1038/s41588-018-0237-2.

Zhang X, Zhang S, Zhao Q, Ming R, Tang H. 2019. Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat Plants. 5(8):833–845. doi:10.1038/s41477-019-0487-8.

Zhou C, McCarthy SA, Durbin R. 2022. YaHS: yet another Hi-C scaffolding tool. Bioinformatics. 39(1):btac808. doi:10.1093/bioinformatics/btac808.

Editor: P. Ingvarsson