G3, 2026, **16(3),** jkaf317
https://doi.org/10.1093/g3journal/jkaf317
Advance Access Publication Date: 3 January 2026
Genome Report

Chromosome-scale genome assembly for yellow wood sorrel, *Oxalis stricta*

Joshua C. Wood [1], John P. Hamilton,[1,2] Brieanne Vaillancourt,[1] Julia Brose,[1,3] Patrick P. Edger,[4] C. Robin Buell [1,2,3,5,6,*]

[1]Center for Applied Genetic Technologies, University of Georgia, Athens, GA 30602, United States
[2]Department of Crop and Soil Sciences, University of Georgia, Athens, GA 30602, United States
[3]Department of Plant Biology, Michigan State University, East Lansing, MI 48824, United States
[4]Department of Horticulture, Michigan State University, East Lansing, MI 48824, United States
[5]Institute of Plant Breeding, Genetics, and Genomics, University of Georgia, Athens, GA 30602, United States
[6]The Plant Center, University of Georgia, Athens, GA 30602, United States

*Corresponding author: Center for Applied Genetic Technologies, University of Georgia, Athens, GA 30602, United States. Email: Robin.Buell@uga.edu

Yellow wood sorrel (*Oxalis stricta* L.), also known as sourgrass, juicy fruit, or sheep weed, is a member of the Oxalidaceae family. Yellow wood sorrel is commonly considered a weed, and while native to North America, it is distributed across Europe, Asia, and Africa. To date, only 2 other genomes from the Oxalidaceae family have been published, star fruit (*Averrhoa carambola* L.) and *Oxalis articulata* Savigny. Here, we present a chromosome-scale assembly for *O. stricta*, revealing its allotetraploid nature and synteny within its 2 subgenomes as well as synteny with *A. carambola* and *O. articulata*. Using Oxford Nanopore Technologies long-read sequences coupled with chromatin capture sequencing, we generated a 436 Mb chromosome-scale assembly of *O. stricta* with a scaffold N50 length of 36.2 Mb that is anchored to 12 chromosomes across the 2 subgenomes. Assessment of the final genome assembly using the Long Terminal Repeat Assembly Index yielded a score of 13.12, and assessment of Benchmarking Universal Single Copy Orthologs revealed 99.6% complete orthologs; both metrics are suggestive of a high-quality reference genome. Total repetitive sequence content in the *O. stricta* genome was 39.7% with retroelements being the largest class of transposable elements. Annotation of protein-coding genes yielded 61,550 high-confidence genes encoding 115,089 gene models. Synteny between the 2 *O. stricta* subgenomes was present in 93 syntenic blocks containing 40,750 genes, of which, 76.47% were present in 1:1 syntenic relationships between the 2 subgenomes. The availability of an annotated chromosome-scale high-quality genome assembly for *O. stricta* will provide a launching point to understand the high fecundity of this weed and provide further foundation for comparative genomics within the Oxalidaceae.

Keywords: wood sorrel; genome assembly; subgenome; synteny

Introduction

Oxalis stricta L., also known as yellow wood sorrel, is a pervasive species in North America and is considered a weed (Holt and Elmore 1985) (Fig. 1). Yellow wood sorrel belongs to the Oxalidaceae, a name that aptly reflects the plants' production of oxalic acid. The Oxalidaceae contains numerous species of botanical interest, including star fruit (*Averrhoa carambola* L.) and oca (*Oxalis tuberosa* Molina), a tuberous food crop in South America, as well as the ornamentals *Oxalis triangularis* A. St.-Hil. (false shamrock) and *Oxalis versicolor* L. (candy cane sorrel). To date, the only species from Oxalidaceae with a genome sequence are star fruit and *Oxalis articulata* Savigny (Wu et al. 2020; Yang et al. 2025).

Yellow wood sorrel boasts edible leaves and seed pods containing oxalic acid and high levels of vitamin C, which impart a strong, tangy flavoring that contributes to its use as an herb (Shad et al. 2013). Despite these potential culinary applications, yellow wood sorrel is widely considered an invasive weed due to several key characteristics: its high fecundity, with each plant producing numerous seeds (Stevens 1932); the ability of its seed capsules to explosively disperse seeds up to 2 m (van der Pijl 2012); the absence of a significant period of seed dormancy allowing for rapid germination; and its production of rhizomes, enabling perennial growth and persistence (Marshall 1987; Christopher Marble 2018).

O. stricta is a tetraploid with a base chromosome number of 6 ($2n = 4x = 24$) with reported 1C genome sizes that range between 528 Mb (Vaio et al. 2013) and 547 Mb (Bai et al. 2012). In this study, we generated a chromosome-scale genome assembly for *O. stricta* comprised of 12 chromosomes, spanning both subgenomes, with a total assembly size of 436 Mb. We annotated 61,550 representative high-confidence genes encoding 115,089 gene models. This genomic resource will facilitate research within the Oxalidaceae, which contain oxalic acid as well as the presence and absence of below ground storage organ formation.

Materials and methods

Genomic DNA isolation and sequencing

Seeds of *O. stricta* were obtained from EdenWilds (Brooklyn, NY USA) and subjected to single seed descent from which cuttings were used from a single S1 individual for genome and

Received on 14 October 2025; accepted on 21 December 2025

Fig. 1. Picture of yellow wood sorrel, *O. stricta*.

transcriptome experiments to ensure genetic uniformity. Plants were grown in a growth chamber under 21.1 °C day/15.6 °C night with 500 µmol/m^2/s light intensity under a 15-h photoperiod. DNA for short-read sequencing was isolated from mature leaves from plants dark-treated for 24 h prior to harvest using the Qiagen DNeasy Plant Pro Kit (Germantown, MD) and a library constructed using the PerkinElmer NEXTFLEX Rapid XP DNA-Seq Kit HT with NEXTFLEX UDI Barcodes (PerkinElmer, Waltham, MA). Sequencing was performed on an Illumina NovaSeq 6000 (San Diego, CA) generating paired end 150 nt reads (Illumina, San Diego, CA; Supplementary Table 1). Library preparation and sequencing were performed by the Texas A&M AgriLife Research: Genomics and Bioinformatics Service.

Plants were dark-treated for 24 h prior to harvesting of leaf tissue for high molecular weight DNA isolation as described previously (Vaillancourt and Buell 2019). High molecular weight DNA was input into the Oxford Nanopore Technologies (ONT) Ligation sequencing gDNA kit (SQK-LSK110) and the resulting libraries sequenced on R9 FLO-MIN106 Rev D flow cells on a MinION (ONT, Oxford, UK; Supplementary Table 1). Young leaves were collected from plants grown in a greenhouse at 25 °C day/15 °C night, with 300 µmol/m^2/s light intensity under 15 h of light. Tissue was then used to construct 2 Hi-C libraries using the Proximo Plant v4.0 protocol with a modified fragmentation enzyme cocktail containing *Dpn*II, *Dde*I, *Hinf*I, and *Mse*I (Phase Genomics, Seattle, WA). Sequencing was performed on an Illumina NovaSeq 6000 at the Michigan State University Research Technology Support Facility generating paired end 150 nt reads (Supplementary Table 1).

RNA isolation and sequencing

For gene expression abundance estimation and genome annotation, we constructed a replicated developmental tissue atlas that included a set of abiotic and biotic stresses to fully capture the *O. stricta* transcriptome. Plants were grown in a growth chamber under 21.1 °C day/15.6 °C night with 500 µmol/m^2/s light intensity under a 15-h photoperiod (6 AM to 9 PM). Core developmental tissues included stem, flower (closed bud, open bud), fibrous root, as well as leaf tissue from a 24-h diurnal time course sampled every 4 h (Supplementary Table 1). To mimic stress conditions, leaves were treated with methyl jasmonate (MeJA, 250 µM) and benzothiadiazole (BTH, 100 µg/mL) via leaf drenches with collection of leaves and roots occurring 24 h after treatment. For salt treatment, 100 mL of 150 mM NaCl solution was applied to the pots and samples (roots, leaves) collected after 24 h. For cold treatment, plants were subjected to a constant temperature of 10 °C for 24 h prior to collection. For heat treatment, plants were watered well to avoid drought stress and the temperature increased to 37 °C (day) and 28 °C (night) and samples (roots, leaves) collected after 24 h. Finally, a drought treatment was conducted with leaves being collected after visible wilting was observed. Except for the time course, all samples were collected at 10 AM, 4 h post-dawn.

RNA was isolated using a modified hot borate method (Wan and Wilkins 1994). Briefly, extraction buffer was heated to 80 °C and added to 100 mg of tissue before vortexing thoroughly. Proteinase K was then added and mixed by inversion before DL-dithiothreitol and additional heated extraction buffer were added. Centrifugation at 10,000 rpm was used to pellet plant debris, the supernatant was collected and mixed with one-third volume 8M LiCl followed by an overnight incubation to precipitate the RNA. RNA was pelleted at 13,000 RPM and washed twice with cold 2 M LiCl and re-precipitated at −20 °C for 50 min using sodium acetate and ethanol. Following precipitation, the RNA was pelleted at 13,000 RPM and washed twice with 70% ethanol. Residual DNA was removed using the TURBO DNase kit (Invitrogen, Waltham, MA). RNA-Seq libraries were constructed using NEXTFLEX Poly(A) Beads 2.0 for PolyA selection followed by the PerkinElmer NEXTFLEX Rapid Directional RNA-Seq Kit 2.0 with NEXTFLEX RNA-Seq 2.0 Unique Dual Index Barcodes (PerkinElmer, Waltham, MA). Sequencing was performed on an Illumina NovaSeq 6000 (Illumina, San Diego, CA) generating paired end 150 nt reads. Library preparation and sequencing were performed by the Texas A&M AgriLife Research: Genomics and Bioinformatics Service. Full-length cDNA libraries were constructed using the ONT PCR-cDNA Barcoding kit (SQK-PCB109) and sequenced on a MinION using R9 FLO-MIN106 Rev D flow cells (ONT, Oxford, UK; Supplementary Table 1).

Genome assembly

Genome heterozygosity was examined using GenomeScope (v2.0) (Ranallo-Benavidez et al. 2020) with the Illumina whole-genome shotgun sequence reads, setting the ploidy to 4, the *k*-mer size to 21, and the max *k*-mer coverage to 300. The ONT genomic data were base-called using Guppy (v5.0.14) (https://nanoporetech.com/software/other/guppy/history?version=5-0-14) using the high accuracy model (dna_r9.4.1_450bps_hac.cfg) and the parameters –trim_strategy dna and –calib_detect. Reads shorter than 10 kb were removed from the read pool using seqkit (v0.16.1) (Shen et al. 2024). Reads 10 kb or greater were input into the Flye (v2.9) assembler (Kolmogorov et al. 2019) with options –nano-raw, –iterations 0, and –scaffold. The assembly was then

polished using 2 rounds of Racon (v1.4.20) (Vaser et al. 2017) with the parameters set as follows: -m 8 -x -6 -g -8 -w 500 -u. After Racon, 2 rounds of Medaka (v1.4.4) (https://github.com/nanoporetech/medaka) polishing were completed using the basecaller model r941_min_hac_g507. Pilon (v1.24) (Walker et al. 2014) was then run using the Illumina whole-genome shotgun reads in 3 consecutive rounds using the options –frags and –fix bases. Pseudomolecules were constructed using the Hi-C libraries as input into Juicer (v1.6) (Durand et al. 2016b) and 3D-DNA (v180922) (Dudchenko et al. 2017) with options -i 10000 and -r 5. Juicebox (Durand et al. 2016b) was used to complete manual curation and the creation of the final chromosomes. The assembly was then filtered to remove contigs less than 10 kb using seqkit (v0.16.1).

Multiple methods were used to assess the quality and completeness of the final assembly. First, the KAT software with the *comp* command (v2.4.2) (Mapleson et al. 2017) was used to assess completeness of the final assembly based on *k*-mer representation in the final assembly vs that in the whole-genome short reads. Second, the completeness of the final genome was assessed by determining the fraction of Benchmarking Single Copy Orthologs (BUSCO) using BUSCO v5.8.2 (Manni et al. 2021) (*Embryophyta* odb 10). Third, the Long Terminal Repeat (LTR) Assembly Index (LAI) metric (Ou et al. 2018) was used to determine contiguity of LTRs. Intact LTR-RTs (LTR retrotransposons) were identified using LTRharvest (v1.6.2) (Ellinghaus et al. 2008), LTR_FINDER_parallel (v1.0.7) (Ou and Jiang 2019), and LTR_retriever (v2.9.0) (Ou and Jiang 2018). Options for LTRharvest were -minlenltr 100 -maxlenltr 7000 -mintsd 4 -maxtsd 6 -motif TGCA -motifmis 1 -similar 85 -vic 10 -seed 20. The options for LTR_FINDER_parallel were -size 1000000 -time 300.

Genome annotation

The genome assembly was repeat masked by first creating a custom repeat library (CRL) for the genome. Repeats were first identified using RepeatModeler (v2.03) (Flynn et al. 2020) with protein-coding genes then removed using ProtExcluder (v1.2) (Campbell et al. 2014) to create a CRL. The CRL was then combined with Viridiplantae repeats from RepBase (v20150807) (Bao et al. 2015) creating the final CRL. The genome assembly was repeat-masked using the final CRL using RepeatMasker (v4.1.2-p1) (Chen 2004) with the parameters -e ncbi -s -nolow -no_is -gff.

RNA-Seq libraries were processed for genome annotation by first cleaning with Cutadapt (v2.10) (Martin 2011) using a minimum length of 100 nt and quality cutoff of 10 then aligning the cleaned reads to the genome assembly using HISAT2 (2.1.0) (Kim et al. 2019). Oxford Nanopore (ONT) cDNA reads were processed with Pychopper (v2.5.0) (https://github.com/epi2me-labs/pychopper), and trimmed reads greater than 500 nt were aligned to the genome assembly using minimap2 (v2.17-r941) (Li 2021) with a maximum intron length of 5,000 nt. The aligned RNA-Seq and ONT cDNA reads were each assembled using Stringtie (v2.2.1) (Kovaka et al. 2019), and transcripts less than 500 nt were removed.

Initial gene models were created using BRAKER2 (v2.1.6) (Brůna et al. 2021) using the soft-masked genome assembly and the aligned RNA-Seq libraries as hints. The gene models were then refined using 2 rounds of PASA2 (v2.5.2) (Campbell et al. 2006) to create a working gene model set. High-confidence gene models were identified from the working gene model set by filtering out gene models lacking expression evidence, or a PFAM domain match, or were a partial gene model, or contained an interior stop codon. A representative gene model was then identified for all high-confidence genes using the model with the longest coding sequence. Functional annotation was assigned by searching the working gene models proteins against the TAIR (v10) (Lamesch et al. 2012) database and the Swiss-Prot plant proteins (release 2015_08) database using BLASTP (v2.12.0) (Altschul et al. 1990) with a *e*-value cutoff of 1e−5 and the PFAM (v35.0) (Mistry et al. 2021) database using PfamScan (v1.6) using the default parameters. The search results were processed in the order of TAIR, PFAM, and Swiss-Prot, and the annotation was assigned by using the first match in the ordered search results. Transposable elements were annotated using Extensive de novo TE Annotator (EDTA) (v2.2.2) (Ou et al. 2019) with the genome and the working gene model CDS sequences as input.

Assigning subgenomes

To define the subgenomes, we employed SubPhaser (v1.2.6) (Jia et al. 2022) with default options, which functions by searching for subgenome-specific *k*-mers to assign the chromosomes to a given subgenome.

Gene expression analyses

RNA-Seq reads were checked for quality with FastQC (v0.11.9) (https://www.bioinformatics.babra-ham.ac.uk/projects/fastqc) and MultiQC (v1.11) (Ewels et al. 2016) before cleaning with Cutadapt (v4.4) using the options -q 30, -m 40, –trim-n, and -n 2. Reads were checked for contaminants using Kraken2 (v2.1) (Wood et al. 2019) with the database k2_pluspfp_20220908. Reads were then subsampled using seqtk (v1.3) (https://github.com/lh3/seqtk.git) to the median number of total reads after cleaning (65 M reads). To determine expression abundances, the quant algorithm of kallisto (v0.48.0) (Bray et al. 2016) was used along with the high-confidence representative gene models with the stranded option set (–rf-stranded) and the *k*-mer size at the default of 31.

Comparative genome analyses

GENESPACE (v1.3.1) (Lovell et al. 2022) was used to identify syntelogs, genes found in colinear blocks on the chromosomes, between *A. carambola* (Wu et al. 2020), *O. articulata* (Yang et al. 2025), and *O. stricta*. The syntenic orthologs, hereafter "syntelogs," were extracted using the "PASS" flag from the *query_pangenes* function. OrthoFinder2 (v2.5.5) (Emms and Kelly 2019) was used to generate orthologous and paralogous groups using the predicted proteomes of *Arabidopsis thaliana* (Cheng et al. 2017), *A. carambola* (Wu et al. 2020), *Cephalotus follicularis* (Fukushima et al. 2017), *Oryza sativa* (Kawahara et al. 2013), *O. articulata* (Yang et al. 2025), *O. stricta*, and *Solanum tuberosum* (Pham et al. 2020). Gene Ontology (GO) associations were generated using InterProScan (v 5.69-101.0) (Jones et al. 2014) and enrichment of GO associations determined using TopGO (2.56.0) (Alexa et al. 2006).

Results and discussion

Assembly of the *O. stricta* genome

A biomodal distribution of *k*-mers was observed using GenomeScope which revealed a heterozygosity of 0.0327% (Supplementary Fig. 1). A total of 44.9 Gb of genomic ONT long reads with an N50 read length of 37.2 kb (~83× coverage) were used to assemble the *O. stricta* genome using the Flye genome assembler software (Supplementary Table 2). Following polishing with ONT long reads and Illumina short reads and scaffolding using Hi-C reads (Fig. 2), the *O. stricta* genome assembly totaled 436,804,292 bp present on 99 scaffolds with an N50 scaffold length

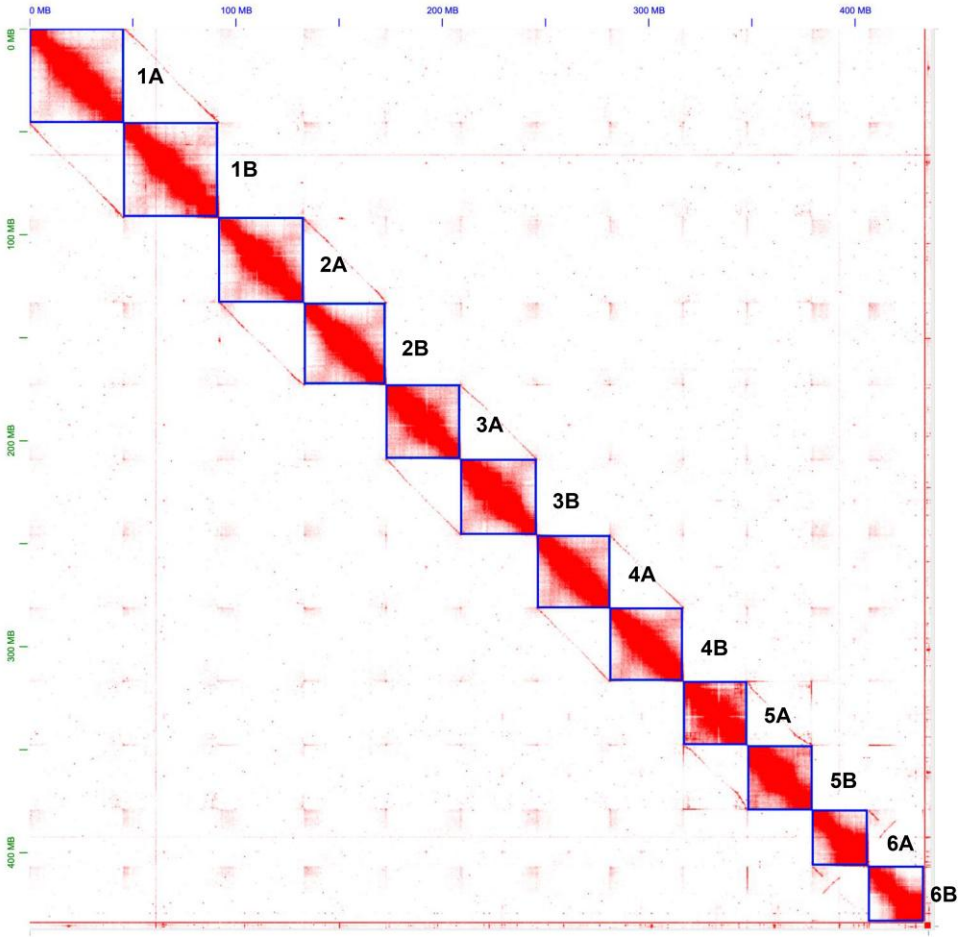

Fig. 2. Hi-C contacts for the *O. stricta* genome visualized by Juicebox 1 (Durand et al. 2016a). Diagonal interactions indicate relationships between blue boxes (pseudomolecules) which show subgenome similarity. Labels denote the 6 chromosomes (1 to 6) per subgenome (A or B).

Table 1. Assembly statistics for the *O. stricta* assembly.

Genome	Number of sequences	Total assembly size (bp)	Min scaffold length (bp)	Max scaffold length (bp)	N50 scaffold length (bp)
O. stricta (subgenome A)	6	216,994,121	27,369,382	46,093,068	36,226,067
O. stricta (subgenome B)	6	216,800,702	26,919,296	46,270,890	36,921,660
O. stricta (v3)	99	436,804,292	10,023	46,270,890	36,226,067

of 36,226,067 bp (Table 1; Supplementary Table 3). Only 36 gaps are present within the assembly, totaling 18,000 bp of sequence. The 12 chromosomes represent 433,794,823 bp of sequence with the remaining 3,009,469 bp of unanchored sequence present on the remaining 87 scaffolds, resulting in a high level of continuity (Supplementary Tables 3 and 4). KAT also revealed substantial representation of *k*-mers as single copy in the final assembly and that the assembly was not missing *k*-mers present in the whole-genome shotgun sequence reads (Supplementary Fig. 2). The LAI assays for intact LTR retrotransposons (LTR-RTs) to evaluate assembly continuity with a higher LAI score indicative of a more complete genome due to a greater number of intact LTR-RTs. The *O. stricta* LAI score was 13.12, placing the assembly within the category of a "reference quality" genome (Ou et al. 2018). Assessment of genome completeness using BUSCO revealed 99.6% complete BUSCOs with a majority of orthologs duplicated suggesting the presence of subgenomes in the assembly (Supplementary Table 5), consistent with the reported tetraploidy of *O. stricta* (Vaio et al. 2013). Overall, these metrics indicate a high-quality, chromosome-scale reference genome assembly for *O. stricta*.

Genome annotation

De novo annotation of repetitive sequences followed by repeat masking revealed 39.7% of the *O. stricta* genome was repetitive with retroelements and unclassified interspersed repeats composing most of the repetitive sequence (Fig. 3; Supplementary Table 6). Annotation of protein-coding genes revealed 61,550 representative high-confidence genes that encoded 115,089 gene models attributable to the extensive RNA-Seq and full-length cDNA sequences available for annotation; 121,961 working gene models were annotated (Supplementary Table 7). The chromosome-anchored subgenomes represent 99.7% of the gene models, with only a small fraction (0.3%) encoded on the

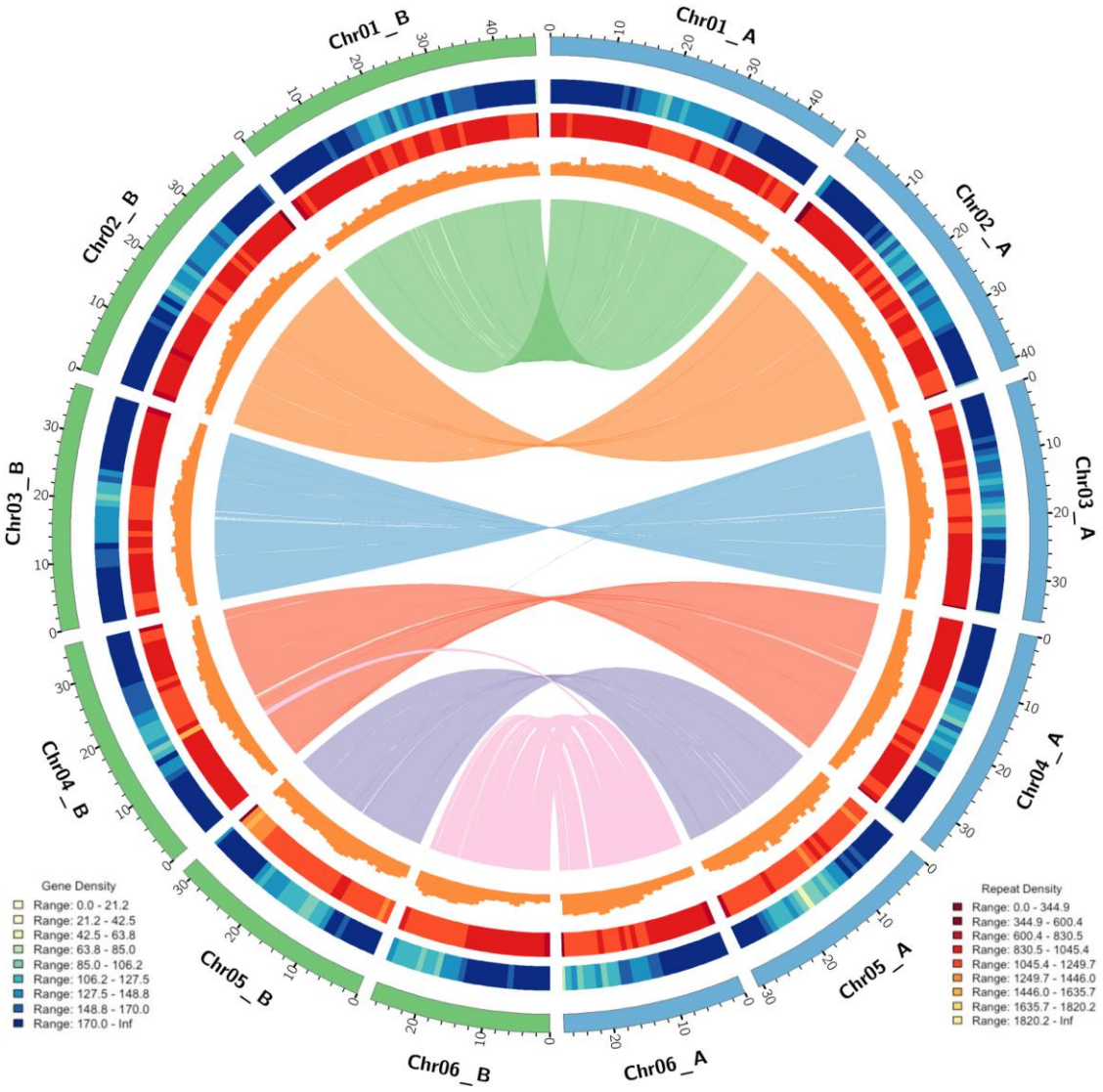

Fig. 3. Circos plot (Krzywinski et al. 2009) of the *O. stricta* genome. Rings (outermost to innermost) represent the 12 chromosomes binned into subgenomes A (blue) or B (green) in Mb, heatmap of gene density, total repeat density heatmap, histogram of transposable element density, and synteny between the 2 subgenomes. Heatmap and histogram bins are 1 Mb in size.

unanchored scaffolds. BUSCO assessment of the representative high-confidence gene model gene annotations closely matched the genome-level BUSCO assessment, with 96.3% complete BUSCOs that were mostly duplicated with very few missing BUSCO orthologs (Supplementary Table 5).

O. stricta is an allotetraploid

Previous reports indicate that *O. stricta* can occur as a tetraploid (Vaio et al. 2013). This is consistent with the pairing of homeologous chromosomes observed in our Hi-C signal, which reveals the presence of 2 subgenomes within our assembly, as demonstrated by diagonal similarities in the contact map (Fig. 2). This was further supported by SubPhaser (Jia et al. 2022) identifying subgenome-specific *k*-mers for subgenomes A (393) and B (409). Examination of the normalized proportions of subgenome-specific *k*-mers across the chromosomes showed mixing of the subgenomes since their allohybridization (Supplementary Fig. 3). Examination of each subgenome separately showed high percentage of single copy BUSCO's with 91.6 and 94.0% for subgenome A and B, respectively (Supplementary Table 5). Assessing the subgenomes independently revealed a mere 3.9 and 3.2% duplicated BUSCO's present in subgenome A and B, respectively. This reduction in duplicated BUSCOs in the full genome vs subgenome level analysis further supports that the subgenome assignment did not place 2 of the same chromosomes within one given subgenome. Distribution of genes across the 2 subgenomes was also even, with subgenome A containing 30,183 representative high-confidence genes and subgenome B containing 31,095 representative high-confidence genes.

Transposable elements in *O. stricta*

To further annotate transposable elements within *O. stricta*, we used the EDTA software revealing 102.6 Mb of transposable element sequences accounting for 23.50% of the assembly (Supplementary Table 8; Fig. 3). Of the different transposable element classes, there were 13.3% (58.1 Mb) and 10% (43.7 Mb) of class I and class II transposable elements, respectively. The class I TEs were comprised of 1.3% (5.8 Mb) LINE elements and

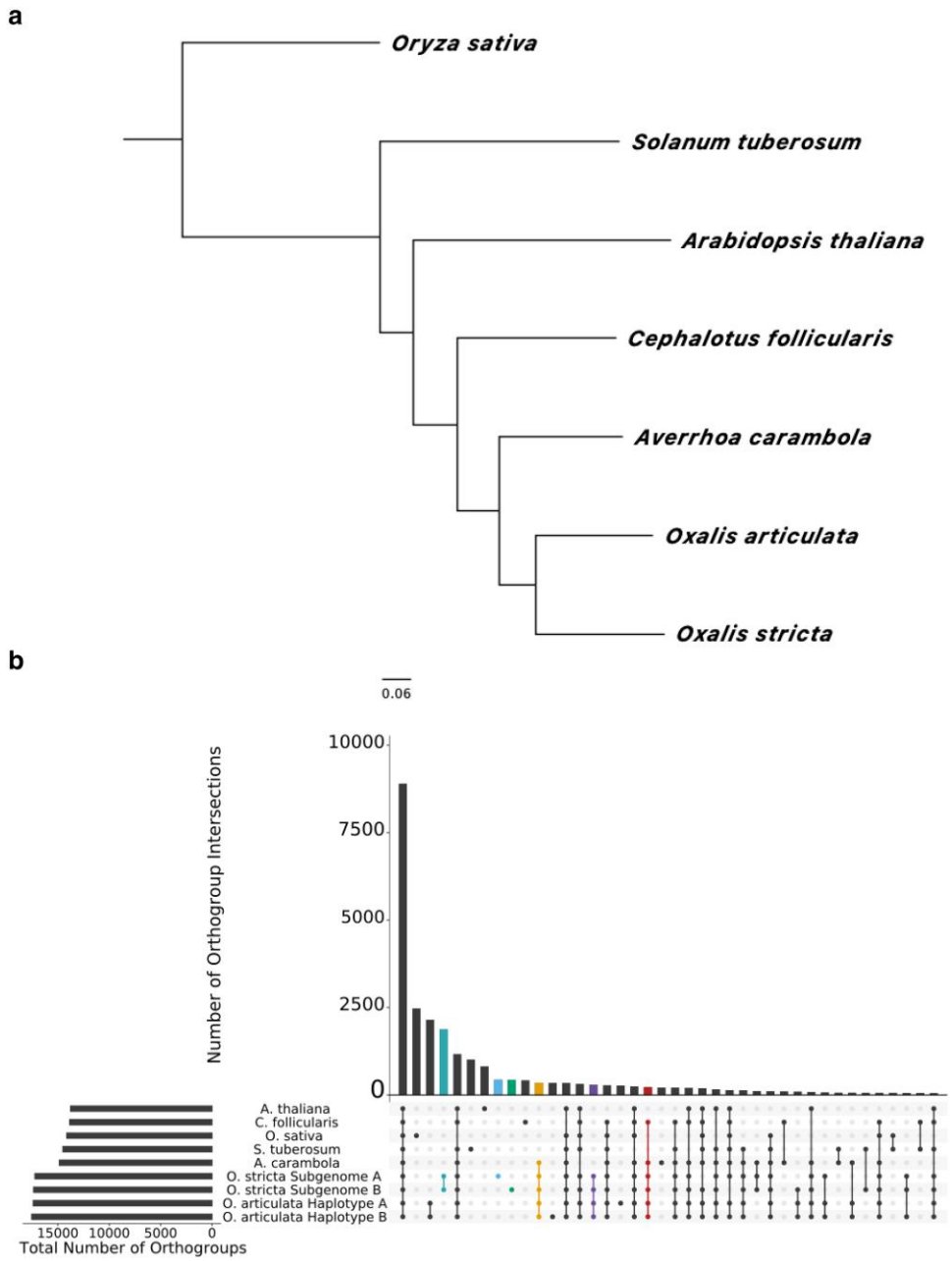

Fig. 4. Phylogeny and orthology of *O. stricta*. a) Phylogeny of *O. stricta* and other angiosperms using Orthofinder2 (Emms and Kelly 2019). b) Upset plot of orthogroups identified between *A. thaliana*, *A. carambola*, *C. follicularis*, *O. sativa*, *O. stricta*, and *S. tuberosum* using Orthofinder2. Light blue represents *O. stricta* subgenome A–specific orthogroups, green represents *O. stricta*–specific subgenome B orthogroups, teal represents *O. stricta* subgenome shared orthogroups, purple represents *Oxalis*-specific orthogroups, orange represents Oxalidaceae-specific orthogroups, and red represents Oxalidales-specific orthogroups.

11.9% (52.2 Mb) LTR elements. The class II TEs were comprised of 6.6% (29 Mb) TIR elements and 3.3% (14.6 Mb) Helitron elements. Transposable elements can differ among subgenomes in an allopolyploid (Hosaka et al. 2024) and subgenome A contained 47.38 Mb (22.06%) of transposable elements while subgenome B contained 42.9 Mb (19.8%) of transposable elements, showing a slight difference between the 2 subgenomes. While differences in individual transposable element classes were minimal between the 2 subgenomes, the 3 largest differences (>0.5% difference between subgenomes) were for *Tc1 Mariner* and *Mutator* elements that were more prevalent in subgenome A. *RTE* elements were present only within subgenome B with 695 elements representing 488 kb.

Orthology analyses reveal lineage-specific genes within *O. stricta*

OrthoFinder2 was used to build a phylogeny for *O. stricta* that was consistent with the known phylogenetic relationships of these species (Fig. 4a). *C. follicularis*, a member of the Oxalidales, was correctly placed sister to the 3 Oxalidaceae species in the phylogeny. The 2 *Oxalis* species were also placed sister to one another. Clustering of 276,074 proteins from these seven species resulted in 27,506 orthogroups. Of these, 223 orthogroups were specific to the Oxalidales, 350 orthogroups were specific to the Oxalidaceae, 291 orthogroups were specific to the *Oxalis* genus and 2,741 orthogroups were unique to *O. stricta* which was comprised of 8,234 genes (Fig. 4b). Within *O. stricta*, there were

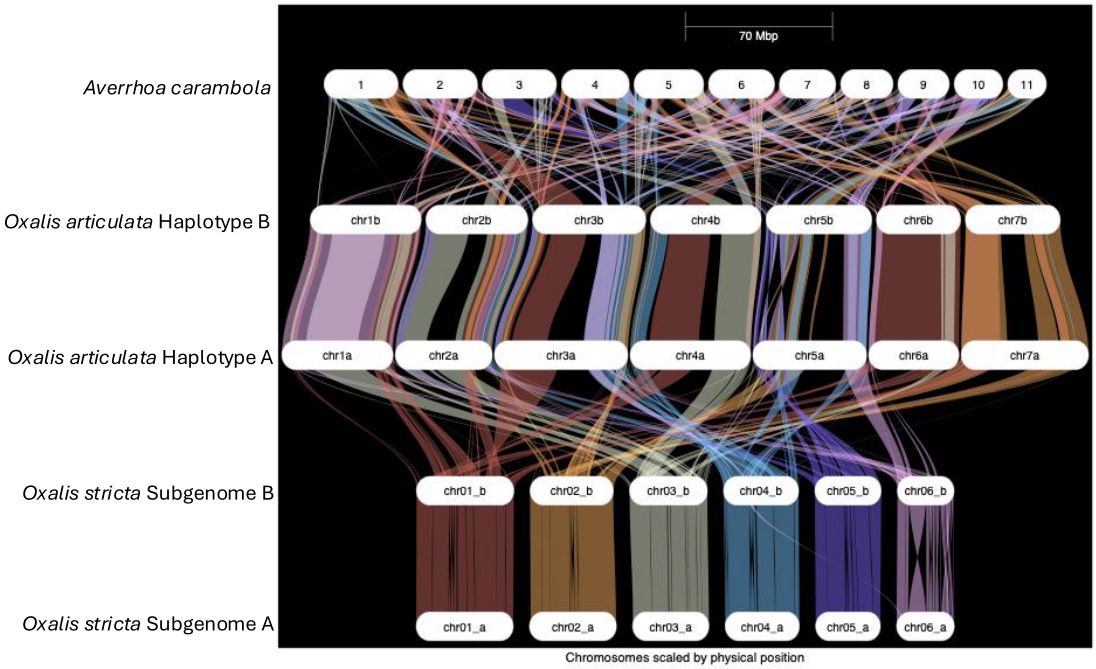

Fig. 5. Synteny between the 2 *O. stricta* subgenomes, the 2 haplotypes of *O. articulata*, and *A. carambola*.

orthogroups specific to each subgenome, with subgenome A containing 435 exclusive orthogroups and subgenome B containing 432 exclusive orthogroups. The subgenome A-exclusive orthogroups contained 1,173 genes while the subgenome B-exclusive orthogroups contained 1,121 genes.

O. stricta subgenomes are highly syntenic

Synteny analysis between *O. stricta* subgenomes A and B revealed 93 blocks containing 40,750 syntenic genes spanning across 211 and 212 Mb in subgenome A and B, respectively (Supplementary Table 9; Fig. 3). Examination of the syntenic relationships revealed that a majority of syntelogs exist within a 1:1 relationship between the 2 subgenomes. A total of 15,581 1:1 syntelogs were present, accounting for 76.47% of the syntenic relationships between the 2 subgenomes. A total of 9,588 other genes had syntenic relationships, with the next most prevalent relationship being 2:2 between the subgenomes. These 2:2 syntelogs comprised 7,284 genes in total and represent instances of paralogous genes. As these paralogs are present in both subgenomes, their duplication is likely ancestral to the most common ancestor of the 2 subgenomes. The next 2 most common relationships within the syntelogs were 1:2 and 2:1, representing 1,197 and 1,047 genes, respectively, that likely represent recent gene duplications or gene losses that uniquely occurred within one of the subgenomes.

Synteny within the Oxalidaceae

We performed synteny analysis between the *O. stricta* subgenomes, *A. carambola*, and the haplotypes of *O. articulata* (Fig. 5). A total of 347 syntenic blocks comprised of 29,138 genes were found between subgenome A and *A. carambola*, with similar synteny between *A. carambola* and *O. stricta* subgenome B observed (Supplementary Table 10). These syntenic genes were dominated by 1:1:1 syntelogs, totaling 10,560 triplets, between *O. stricta* subgenome A, *O. stricta* subgenome B and *A. carambola*. When comparing the subgenomes of *O. stricta* with the haplotypes of *O. articulata*, we found that there were more syntenic blocks and collinear gene content with haplotype A of *O. articulata* for both subgenomes of *O. stricta*. Surprisingly, the percentage of collinear genes between *O. stricta* and *O. articulata* was less than that between *O. stricta* and *A. carambola*. The *Oxalis* genus is large, spanning over 500 species (Azkue 2000), of which *O. stricta* and *O. articulata* belong to different sections within the genus (Oberlander et al. 2009). This diversity may explain why *O. stricta* has more gene collinearity with star fruit than *O. articulata*; however, differences in assembly and annotation make this comparison difficult. Synteny across all available Oxalidaceae genomes/subgenomes revealed 8,999 genes present in a 1:1:1:1:1 relationship.

Subgenome dominance in *O. stricta*

The 15,581 1:1 syntelogs between the 2 *O. stricta* subgenomes were examined for subgenome dominance using gene expression data from 13 replicated RNA-Seq datasets comprising different tissues and/or treatments (Supplementary Tables 1, 11, and 12). Of these 1:1 syntelogs, a total of 29,991 of the genes had a transcript per million (TPM) greater than 0 in our expression profiling datasets. A Wilcoxon rank sum test of the mean log(TPM + 1) values was conducted to test whether syntelogs within a given subgenome showed dominance in certain tissues or conditions, excluding the diurnal time course due to it containing substantially more samples than the other tissue and treatment sets. Across the 12 datasets included, half showed bias, including both flower (open and closed), root (control and salt treated), leaf cold treatment, and stem tissues, in which the median TPMs are higher in subgenome B than subgenome A (Fig. 6; Supplementary Fig. 4). Syntelog pairs were defined as biased if the log2 expression fold difference between the subgenomes was greater than |2| (Supplementary Fig. 5). Examination of the paired syntelog expression showed that the most significant bias for subgenome B was within the flower tissues, with a total of 2,263 pairs in closed flowers and 2,172 pairs in open flowers (Supplementary Table 12). Further examination of the paired syntelog expression within the other tissues revealed additional bias for subgenome B with 2,159 pairs for root, 2,274 pairs for salt treated roots, 2,251 pairs for leaf cold

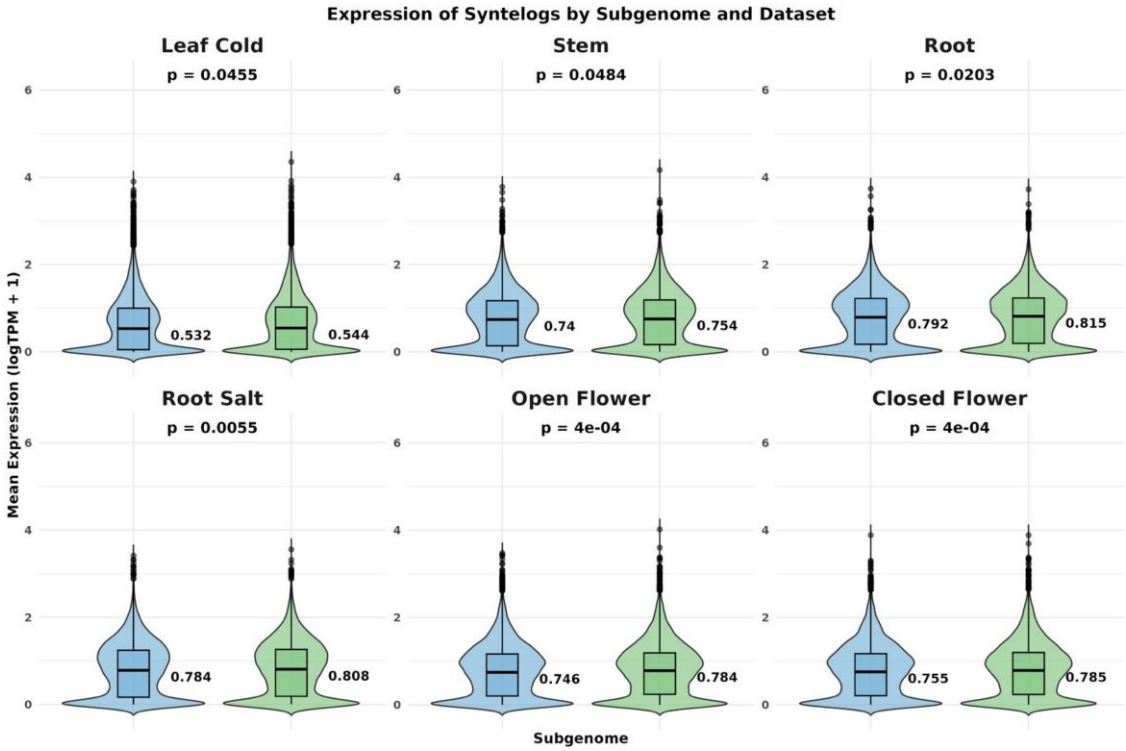

Fig. 6. Expression of 1:1 syntelogs in *O. stricta* and their expression bias. Violin plots depicting expression of 1:1 *O. stricta* subgenome syntelogs in various tissues/treatments. *P*-values above each subgenome comparison are the pairwise Wilcoxon test assessing for significant differences between the subgenomes. Values beside boxplots are median log(TPM + 1) expression values for each subgenome/dataset. Subgenome A in blue and subgenome B in green.

treatment, and 2,179 pairs for stem (Supplementary Tables 13 to 15).

Subgenome dominance, a phenomenon in which one parental subgenome in a polyploid exhibits greater gene expression abundance than other subgenome(s), is hypothesized to be influenced by the degree of similarity between the constituent subgenomes (Garsmeur et al. 2014). Previous studies suggest that subgenome dominance is less pronounced or prevalent in autopolyploids (derived from the duplication of a single genome) or in allopolyploids formed from species with highly similar subgenomes (Garsmeur et al. 2014; Zhao et al. 2017; Alger and Edger 2020; Fang et al. 2024). This is thought to be because the greater the similarity between subgenomes, the fewer genetic conflicts may arise from their co-existence in a single nucleus, thus reducing the need for differential regulation or gene silencing. In our *O. stricta* assembly, the 2 subgenomes exhibit a high degree of similarity, with over 60% of their genes being collinear, indicating a close evolutionary relationship. Consequently, the general lack of significant subgenome dominance observed across 6 of our 12 datasets aligns with this hypothesis, suggesting that the close relatedness of the *O. stricta* subgenomes has limited the establishment of strong expression biases.

Conclusions

Here, we present a chromosome-scale genome assembly for allotetraploid yellow wood sorrel (*O. stricta*). A comparative analysis of the 2 subgenomes revealed the presence of large syntenic blocks and a high degree of collinearity, with over 60% of the genes occupying corresponding positions. Furthermore, synteny analysis comparing the yellow wood sorrel subgenomes with the only 2 other available genomes within the Oxalidaceae, star fruit (*A. carambola*) and *O. articulata*, also showed high percentages (50% and 47%, respectively) of collinear genes, highlighting conserved genomic regions within the family. Subgenome dominance analysis between syntenic gene pairs revealed some tissue-specific expression bias, with flower and root tissues exhibiting favored expression for subgenome B, while the remaining tissues and treatments showed no significant bias in expression between the 2 subgenomes. This newly generated genome assembly represents only the third publicly available genome in the Oxalidaceae, and second within the large *Oxalis* genus, and will provide a valuable resource for future research into the evolutionary and functional genomics of this diverse plant group.

Data availability

Raw sequence data have been deposited in the National Center for Biotechnology Sequence Read Archive under BioProject PRJNA856298. This *O. stricta* whole-genome shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JBLRZZ000000000. The version described in this paper is version JBLRZZ010000000. Large data files (genome assembly, genome annotation, gene expression abundances, EDTA TE annotations) are available in Figshare under doi (https://doi.org/10.6084/m9.figshare.29403401). Scripts are available via this public GitHub repository: https://github.com/woodjosh7/Oxalis_Stricta_Genome.

Supplemental material available at G3 online.

Acknowledgments

We thank the Michigan State University Research Technology Support Facility and the Texas A&M AgriLife Research:

Genomics and Bioinformatics Service for providing sequencing services.

Funding

Funding for this work was provided by an award from the United States National Science Foundation (IOS-2140176 to C.R.B. and P.P.E.), the University of Georgia (C.R.B.), Georgia Research Alliance (C.R.B.), and Georgia Seed Development (C.R.B.).

Conflicts of interest

None declared.

Author contributions

C.R.B. and P.P.E. designed the study. P.P.E. obtained germplasm. J.B. and J.C.W. generated data. B.V., J.P.H., and J.C.W. performed data analyses. J.C.W. and C.R.B. wrote the manuscript; all coauthors reviewed and edited the manuscript.

Literature cited

Alexa A, Rahnenführer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics. 22:1600–1607. https://doi.org/10.1093/bioinformatics/btl140.

Alger EI, Edger PP. 2020. One subgenome to rule them all: underlying mechanisms of subgenome dominance. Curr Opin Plant Biol. 54:108–113. https://doi.org/10.1016/j.pbi.2020.03.004.

Altschul SF, Gish W, Miller W, Meyers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.

Azkue DDE. 2000. Chromosome diversity of South American *Oxalis* (oxalidaceae). Bot J Linn Soc. 132:143–152. https://doi.org/10.1111/j.1095-8339.2000.tb01210.x.

Bai C, Alverson W, Follansbee A, Waller D. 2012. New reports of nuclear DNA content for 407 vascular plant taxa from the United States. Ann Bot. 110:1623–1629. https://doi.org/10.1093/aob/mcs222.

Bao W, Kojima KK, Kohany O. 2015. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 6:11. https://doi.org/10.1186/s13100-015-0041-9.

Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 34:525–527. https://doi.org/10.1038/nbt.3519.

Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. NAR Genom Bioinform. 3:lqaa108. https://doi.org/10.1093/nargab/lqaa108.

Campbell MA, Haas BJ, Hamilton JP, Mount SM, Buell CR. 2006. Comprehensive analysis of alternative splicing in rice and comparative analyses with Arabidopsis. BMC Genomics. 7:327. https://doi.org/10.1186/1471-2164-7-327.

Campbell MS *et al.* 2014. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol. 164:513–524. https://doi.org/10.1104/pp.113.230144.

Chen N. 2004. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics. Chapter 4:Unit 4 10. https://doi.org/10.1002/0471250953.bi0410s05.

Cheng C-Y *et al.* 2017. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. Plant J. 89:789–804. https://doi.org/10.1111/tpj.13415.

Christopher Marble S. 2018. Native weedy pests of the deep south. HortScience. 53:1244–1249. https://doi.org/10.21273/HORTSCI13112-18.

Dudchenko O *et al.* 2017. *De novo* assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 356:92–95. https://doi.org/10.1126/science.aal3327.

Durand NC *et al.* 2016a. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 3:99–101. https://doi.org/10.1016/j.cels.2015.07.012.

Durand NC *et al.* 2016b. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 3:95–98. https://doi.org/10.1016/j.cels.2016.07.002.

Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics. 9:18. https://doi.org/10.1186/1471-2105-9-18.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20:238. https://doi.org/10.1186/s13059-019-1832-y.

Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 32:3047–3048. https://doi.org/10.1093/bioinformatics/btw354.

Fang C *et al.* 2024. Dynamics of accessible chromatin regions and subgenome dominance in octoploid strawberry. Nat Commun. 15:2491. https://doi.org/10.1038/s41467-024-46861-0.

Flynn JM *et al.* 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 117:9451–9457. https://doi.org/10.1073/pnas.1921046117.

Fukushima K *et al.* 2017. Genome of the pitcher plant Cephalotus reveals genetic changes associated with carnivory. Nat Ecol Evol. 1:59. https://doi.org/10.1038/s41559-016-0059.

Garsmeur O *et al.* 2014. Two evolutionarily distinct classes of paleopolyploidy. Mol Biol Evol. 31:448–454. https://doi.org/10.1093/molbev/mst230.

Holt J, Elmore C. 1985. *Oxalis*—biology and control. Ornamentals Northwest Archives. 9:6–8.

Hosaka AJ, Sanetomo R, Hosaka K. 2024. Allotetraploid nature of a wild potato species, *Solanum stoloniferum* schlechtd. et Bché., as revealed by whole-genome sequencing. Plant J. Plant J. 121:e17158. https://doi.org/10.1111/tpj.17158.

Jia K-H *et al.* 2022. SubPhaser: a robust allopolyploid subgenome phasing method based on subgenome-specific k-mers. New Phytol. 235:801–809. https://doi.org/10.1111/nph.18173.

Jones P *et al.* 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics. 30:1236–1240. https://doi.org/10.1093/bioinformatics/btu031.

Kawahara Y *et al.* 2013. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice. 6:4. https://doi.org/10.1186/1939-8433-6-4.

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

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 37:540–546. https://doi.org/10.1038/s41587-019-0072-8.

Kovaka S *et al.* 2019. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biol. 20:278. https://doi.org/10.1186/s13059-019-1910-1.

Krzywinski M et al. 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19:1639–1645. https://doi.org/10.1101/gr.092759.109.

Lamesch P et al. 2012. The Arabidopsis information resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 40: D1202–D1210. https://doi.org/10.1093/nar/gkr1090.

Li H. 2021. New strategies to improve minimap2 alignment accuracy. Bioinformatics. 37:4572–4574. https://doi.org/10.1093/bioinformatics/btab705.

Lovell JT et al. 2022. GENESPACE tracks regions of interest and gene copy number variation across multiple genomes. Elife. 11: e78526. https://doi.org/10.7554/eLife.78526.

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. 38:4647–4654. https://doi.org/10.1093/molbev/msab199.

Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. 2017. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics. 33:574–576. https://doi.org/10.1093/bioinformatics/btw663.

Marshall G. 1987. A review of the biology and control of selected weed species in the genus *Oxalis*: *O. stricta* L., *O. latifolia* H.B.K. and *O. pes-caprae* L. Crop Protection. 6:355–364. https://doi.org/10.1016/0261-2194(87)90068-8.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 17:10–12. https://doi.org/10.14806/ej.17.1.200.

Mistry J et al. 2021. Pfam: the protein families database in 2021. Nucleic Acids Res. 49:D412–D419. https://doi.org/10.1093/nar/gkaa913.

Oberlander KC, Emshwiller E, Bellstedt DU, Dreyer LL. 2009. A model of bulb evolution in the eudicot genus *Oxalis* (Oxalidaceae). Mol Phylogenet Evol. 51:54–63. https://doi.org/10.1016/j.ympev.2008.11.022.

Ou S, Chen J, Jiang N. 2018. Assessing genome assembly quality using the LTR assembly Index (LAI). Nucleic Acids Res. 46:e126. https://doi.org/10.1093/nar/gky730.

Ou S, Jiang N. 2018. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. 176:1410–1422. https://doi.org/10.1104/pp.17.01310.

Ou S, Jiang N. 2019. LTR_FINDER_parallel: parallelization of LTR_FINDER enabling rapid identification of long terminal repeat retrotransposons. Mob DNA. 10:48. https://doi.org/10.1186/s13100-019-0193-0.

Ou S et al. 2019. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 20:275. https://doi.org/10.1186/s13059-019-1905-y.

Pham GM et al. 2020. Construction of a chromosome-scale long-read reference genome assembly for potato. Gigascience. 9:giaa100. https://doi.org/10.1093/gigascience/giaa100.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. Nat Commun. 11:1432. https://doi.org/10.1038/s41467-020-14998-3.

Shad AA, Shah HU, Bakht J. 2013. Ethnobotanical assessment and nutritive potential of wild food plants. J Anim Plant Sci. 23:92–99.

Shen W, Sipos B, Zhao L. 2024. SeqKit2: a Swiss army knife for sequence and alignment processing. Imeta. 3:e191. https://doi.org/10.1002/imt2.191.

Stevens OA. 1932. The number and weight of seeds produced by weeds. Am J Bot. 19:784. https://doi.org/10.1002/j.1537-2197.1932.tb08859.x.

Vaillancourt B, Buell CR. 2019 Sep 26. High molecular weight DNA isolation method from diverse plant species for use with Oxford Nanopore sequencing [preprint]. bioRxiv 783159. https://doi.org/10.1101/783159

Vaio M, Gardner A, Emshwiller E, Guerra M. 2013. Molecular phylogeny and chromosome evolution among the creeping herbaceous *Oxalis* species of sections Corniculatae and Ripariae (Oxalidaceae). Mol Phylogenet Evol. 68:199–211. https://doi.org/10.1016/j.ympev.2013.03.019.

van der Pijl L. 2012. Principles of dispersal in higher plants. Springer.

Vaser R, Sovic I, Nagarajan N, Sikic M. 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 27:737–746. https://doi.org/10.1101/gr.214270.116.

Walker BJ et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 9:e112963. https://doi.org/10.1371/journal.pone.0112963.

Wan CY, Wilkins TA. 1994. A modified hot borate method significantly enhances the yield of high-quality RNA from cotton (*Gossypium hirsutum* L.). Anal Biochem. 223:7–12. https://doi.org/10.1006/abio.1994.1538.

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

Wu S et al. 2020. The genome sequence of star fruit (Averrhoa carambola). Hortic Res. 7:95. https://doi.org/10.1038/s41438-020-0307-3.

Yang W et al. 2025. A haplotype-resolved chromosomal-level genome assembly of Oxalis articulata. Sci Data. 12:856. https://doi.org/10.1038/s41597-025-05237-9.

Zhao M, Zhang B, Lisch D, Ma J. 2017. Patterns and consequences of subgenome differentiation provide insights into the nature of paleopolyploidy in plants. Plant Cell. 29:2974–2994. https://doi.org/10.1105/tpc.17.00595.

Editor: N. Whiteman