bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Genome assembly, annotation and comparative analysis of the cattail *Typha latifolia*

Shane D. Widanagama[1], Joanna R. Freeland[2], Xinwei Xu[3], Aaron B.A. Shafer[4]

[1]Department of Computer Science, Trent University, Peterborough ON, CA, K9L 0G2,

[2]Department of Biology, Trent University, Peterborough ON, CA, K9L 0G2,

[3]Department of Ecology, College of Life Sciences, Wuhan University, Wuhan, China

[4]Department of Forensic Sciences, Trent University, Peterborough ON, CA, K9L 0G2

Genome assembly of *Typha latifolia*

Keywords: PacBio long-read sequencing, Illumina short-read, de novo, Typhaceae, bulrush, broadleaf cattail, hybrids

Corresponding author: Aaron Shafer, Trent University DNA Building, 2140 East Bank Drive, Peterborough, ON, CA, 705 748-1011 x6612, aaronshafer@trentu.ca

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

ABSTRACT

Cattails (*Typha* species) comprise a genus of emergent wetland plants with a global distribution. *Typha latifolia and T. angustifolia* are two of the most widespread species, and in areas of sympatry can interbreed to produce the hybrid *Typha* x *glauca*. In some regions the relatively high fitness of *T.* x glauca allows it to outcompete and displace both parent species, while simultaneously reducing plant and invertebrate biodiversity, and modifying nutrient and water cycling. We generated a high-quality whole genome assembly of *T. latifolia* using PacBio long-read and high coverage Illumina sequences that will facilitate evolutionary and ecological studies in this hybrid zone. Genome size was 287 Mb and consisted of 1,189 scaffolds, with an N50 of 8.706 Mb; 43.84% of the genome were identified as repetitive elements. The assembly has a BUSCO score of 96.03%, and 27,432 genes and 2,700 RNA sequences were putatively identified. Comparative analysis detected over 9,000 shared orthologues with related taxa and phylogenomic analysis supporting *Typha latifolia* as a divergent lineage within Poales. This high-quality scaffold-level reference genome will provide a useful resource for future population genomic analyses and improve our understanding of *Typha* hybrid dynamics.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

INTRODUCTION

Cattails (*Typha* spp.) are aquatic macrophytes that are essential components of wetlands around the world (reviewed in Bansal et al., 2019). These macrophytes often dominate ecosystems through a combination of rapid growth, large size, and sexual and asexual reproduction (Yeo, 1964; Andrews and Pratt, 1978; Miller and Fujii, 2010). Cattails are vital to many wetlands where they cycle nutrients, provide habitat, and aid in bioremediation (Grosshans, 2014; Svedarsky et al., 2016; Bonanno and Cirelli, 2017). However, cattails can also dominate wetlands and in recent decades invasive cattails have been identified in numerous regions, often following anthropogenic changes that have altered water cycles and increased nutrient loads in wetlands (reviewed in Zedler and Kercher, 2004; Bansal et al., 2019). Invasive cattails often form monotypic stands with deleterious effects on local plants and animals, wetland water cycling, and biogeochemical cycles (Yeo, 1964; Grace and Harrison, 1986; Farrer and Goldberg, 2009; Lawrence et al., 2016).

Typha latifolia is likely the most widespread cattail species, occurring on every continent except Antarctica (Smith, 1987). *Typha angustifolia* is also widespread throughout the temperate northern hemisphere (Grace and Harrison, 1986), including North America where it was likely introduced from Europe several centuries ago (Ciotir et al., 2013, 2017). In some regions of sympatry *T. latifolia* and *T. angustifolia* interbreed to produce the hybrid *Typha* x *glauca* (Grace & Harrison, 1986; Ciotir et al., 2017; Bansal et al., 2019). In regions surrounding the Laurentian Great Lakes and St. Lawrence Seaway in North America, *T.* x *glauca* exhibits heterosis (Bunbury-Blanchette et al., 2015; Zapfe and Freeland, 2015), and is more abundant than its parental species

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

(Kirk et al., 2011; Freeland et al., 2013; Pieper et al., 2020). In addition to displacing both parental species, invasive *T.* x *glauca* reduces native plant and invertebrate biodiversity (Lawrence et al., 2016; Tuchman et al., 2009), and alters nutrient cycling and community structure in wetlands (Tuchman et al., 2009; Larkin et al., 2012; Geddes et al., 2014; Lishawa et al., 2014; Lawrence et al., 2017).

Although dominant in some regions of North America, *T.* x *glauca* is uncommon in other regions where the parental species are sympatric, including Europe (Ciotir et al., 2017), eastern Canada (Freeland et al., 2013), and China (Zhou et al., 2016). It is not well understood why hybrids are dominant in some regions but not others, although Tisshaw et al., (2020) suggested that hybrids may be limited in coastal wetlands because they have difficulty germinating in salt-rich environments. In addition, although advanced-generation and back-crossed hybrids have been experimentally generated (Pieper et al., 2017) and identified in natural populations (e.g. Kirk et al., 2011; Freeland et al., 2013; Pieper et al., 2020), it has not been possible to differentiate advanced-generation and backcrossed hybrids based on the small number of species-specific molecular markers that are currently available (Kirk et al., 2011; Snow et al., 2010). Morphology-based assessments are also unreliable due to overlapping phenotypes (Tangen et al., 2021).

A genome-wide suite of SNPs specific to one or the other parent species would greatly facilitate investigations into the dynamics of the *T.* x *glauca* hybrid zone, which is now expanding westwards from the Great Lakes region into the Prairie Pothole Region of Canada and the USA (Tangen et al., 2021). Genome characterization would also facilitate investigations into local adaptation, introgression, and hybrid dynamics;

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

this in turn may inform future management of *T. latifolia,* which is predicted to experience a dramatically reduced distribution following climate change (Xu et al., 2013), likely to the detriment of wetlands throughout its widespread native distribution. Here we report the genome assembly, annotation and analysis of *T. latifolia,* the first fully sequenced species in the family Typhaceae.

MATERIALS & METHODS

Sampling and Sequencing. Leaves of a known *T. latifolia* plant (Figure 1) were taken from an individual grown at Trent University's (Ontario, Canada) greenhouse and ground in liquid nitrogen. DNA was immediately extracted using an E.Z.N.A Plant DNA Kit (Omega Bio-Tek, Inc. Georgia USA) following the manufacturer's instructions for frozen material and eluted in a final volume of 100 µl. DNA quality was assessed on a Tapestation (Agilent Technologies, California USA) before being sent to The Centre for Applied Genomics, Toronto, Ontario, for sequencing. Paired-end reads were sequenced on a single lane of an Illumina HiSeqX system (Illumina, Inc. California USA). Long-reads (LR) were sequenced on one SMRT HiFi cell using a PacBio Sequel II system (Pacific Biosciences of California, Inc. California USA).

Genome Assembly. Adapters and low-quality bases were trimmed from paired-end short reads (SR) using Trimmomatic -v0.36 (Bolger et al., 2014). The optimal k-mer and genome size from the short reads was estimated using Kmergenie -v1.7051 (Chikhi & Medvedev, 2014). We then applied a four-step hybrid assembly: 1) SR assemblies were generated using multiple pipelines: ABySS –v2.2.4 (Jackman et al., 2017) with default settings, 32 threads, a 125 GB total memory; SOAPdenovo2 -r240 (Luo et al., 2012) using default settings, 16 threads, 400 GB of total memory; and Platanus -v1.2.4

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

(Kajitani et al., 2014) using default settings, 32 threads, a 343 GB memory limit, a total memory of 375 GB. We also downsampled the SR data to 20% because really high coverage can be problematic for de Bruijn graphs (Richards & Murali, 2015). The SR assembly with highest quality and completeness was selected by calculating contig N50, total contig length as a percentage of estimated genome size, and the number and percentage of contigs longer than 50 Kb. 2) SR contigs were then aligned to raw long-reads (LR) to create scaffolds using DBG2OLC -v20160205 (options: k 17, AdaptiveTh 0.001, KmerCovTh 2, MinOverlap 20, RemoveChimera 1) (Ye et al., 2016). 3) LR contigs were assembled using Canu –v2.0 (Koren et al., 2017) using the expected genome size inferred from Kmergenie, a maximum memory of 124 GB, and a maximum of 32 threads. And 4) the hybrid assembly (step 2) was merged with Canu LR contigs (step 3) using QuickMerge –v0.3.0 (Chakraborty et al., 2016) following the 2-step strategy described by Solares et al., (2018). Pilon –v1.23.0 (Walker et al., 2014) was used to polish the genome by mapping the Illumina SR back to the genome, thereby correcting base errors and small misassemblies.

Genome Evaluation and Annotation. The *T. latifolia* hybrid genome was assessed for completeness using Benchmarking Universal Single-copy Orthologs (BUSCO) – v3.0.2 (Simão et al., 2015). BUSCOs of the lineage Liliopsida were assessed from OrthoDB release 10 (Waterhouse et al., 2013). We aligned available chloroplast and mitochondrial genomes (Table S1) to our assembly using NUCmer -v3.23 (Kurtz et al., 2004) to identify plastid genomes; scaffolds with long matches (>5000 bp) were extracted and further validated on NCBI Blast. Structural and functional annotation was done in the GenSAS -v6.0 annotation pipeline (Humann et al., 2019). We masked

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

interspersed and simple repetitive elements throughout the genome using a database developed through repeat modeler -v2.0.1 (Smit & Hubley, 2015) in conjunction with repeat masker -v4.0.7 with the NCBI/rmblast search engine, and quick sensitivity (Smit, Hubley & Green, 2015). The hardmasked version of the genome sequence was used for feature prediction. Available RNA-seq data from *T. angustifolia* (SRR15541138) were mapped to the genome using hisat2 -v2.1.0 (Kim et al., 2019) and gene prediction with the RNA-seq alignments was performed by the BRAKER2 -v2.1.1 pipeline, which uses Augustus and GeneMark-ET (Lomsadze et al., 2014, Brůna et al., 2021). ncRNA was predicted using tRNAscan-SE -v2.0 (Lowe & Eddy 1997) and Infernal -v1.1.3 (Nawrocki & Eddy 2013). Refinement of the official gene set was performed with an available *T. latifolia* transcriptome (Moscou, 2017) using PASA -v2.3.3 (Haas et al., 2003).

Comparative Genomics. Single-copy orthologous genes from five other species were identified using OrthoVenn2 with an e-value cutoff of 1e-5, an inflation value of 1.5, and the annotation, protein similarity network, and cluster relationship network enabled (Xu et al., 2019). OrthoVenn2 performs all-against-all genome-wide protein comparisons and groups genes into clusters with the Markov Clustering Algorithm, where a cluster is made up of orthologs and paralogs (Wang et al., 2015). Here we selected *Oryza sativa Japonica, Brachypodium distachyon,* and *Sorghum bicolor* from the family Poaceae*,* and *Ananas comosus* from the family Bromeliaceae. *Arabidopsis thaliana* from the Brassicaceae family was chosen as the dicot outgroup. All proteomes were from the Ensembl database release 104 (Howe et al., 2021). The identified orthologous sequences were then aligned using MAFFT -v7.741 (Katoh & Standley, 2013) and concatenated into single sequences by species using SeqKit -v0.15.0 (Shen et

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

al., 2016). Maximum likelihood phylogenetic analysis was performed using RAxML - v8.2.12 (Stamatakis, 2014) with the PROTGAMMAAUTO substitution model (Darriba et al., 2012). The phylogenetic tree divergence times were estimated using MCMCTree from the PAML package -v4.9j (Yang 1997, 2007), and was calibrated using the divergence time between *Sorghum* and *Oryza* (42-52 Mya) (Kumar et al., 2017).

RESULTS

Sequencing data and genome assembly. A total of 138.6 Gb of raw 151 bp Illumina reads were sequenced (Table S2). Quality filtering and trimming removed 6.5 Gb of sequence data. A recommended k-mer size of 101 bp and estimated genome size of 257 Mb was estimated from the short-read data (Figure S1). Long-read data from the Pacbio Sequel II generated 86.8 Gb of raw data: this included 7,244,218 subreads with a mean length of 11,978.2 bp. All sequencing reads were deposited in the NCBI Sequence Read Archive (Accession No PRJNA751759). The ABySS assembly that used 100% of Illumina reads had the most contiguous genome and was used for the assembly (Table S3). This AbySS SR assembly had a N50 of 0.011 Mb, 365,565 contigs, and 362 contigs longer than 50 Kb (Table S3). The DBG2OLC assemblies (ABySS contigs + raw long reads) improved the assembly statistics but resulted in a smaller than expected genome (Table 1); here we calculated N50 of 0.132 Mb, 1,840 contigs, and 1,445 contigs longer than 50 Kb (Table S4). The LR Canu assembly had an N50 of 8.706 Mb, and contained 1,189 contigs, 821 of which were longer than 50 Kb (95.54%) (Table 1). The final merged and polished hybrid genome assembly – DGB2OLC assembly 1 + Canu - was 287 MB, with an N50 of 8.706 Mb, and consisted of 1,189 scaffolds (Table 1). The GC content was

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

175 38.11% (Table 1). The polished genome has been deposited in the NCBI genome
176 database.

177 **Assembly quality and Annotation.** Approximately 96.03% of BUSCOs were
178 identified (3,148 of 3,278) in the assembly. Of these, 3,025 BUSCOs were complete
179 (92.28%), and 123 were fragmented. Of the complete BUSCOs, 2,461 were single-copies,
180 and 564 had duplicates. Repeats represented 43.84% of the genome (see breakdown in
181 Figure 2). The chloroplast genome was detected and split on two scaffolds, while the
182 entire mitochondrial genome was assembled (Table S5). Total repeat content of the
183 genome is consistent with other Poales genomes (Kawahara et al., 2013, Ming et al.,
184 2015, Redwan et al., 2016): long interspersed nuclear elements (LINEs) comprised
185 1.22% of the genome, while short interspersed nuclear elements (SINEs) were not
186 detected in the genome (Table 2). The annotation pipeline produced 27,432 genes,
187 which coded for 34,911 proteins and 34,974 mRNA sequences. 2,095 rRNA, 502 tRNA,
188 and 214 miRNA sequences were putatively identified, which are similar to sequence
189 counts in related plants (Table S6). No snRNA were identified.

190 **Comparative Genomic Analyses.** Ortholog clustering analysis revealed 9,168 gene
191 families were shared among *T. latifolia, A. comosus, A. thaliana, B. distachyon, O.*
192 *sativa, and S. bicolor* (Figure 2). 1,806 gene families were found to be unique to *Typha*
193 *latifolia* (Figure 2). We aligned 1,900 single-copy gene clusters to conduct phylogenomic
194 analysis (Figure 3). The phylogenomic tree generated supports a divergent *Typha*
195 position in Poales (Figure 3), with *Typha* forming a separate clade with pineapples
196 (*Ananas comosus*). We estimated the bromeliad lineage (*Typha-Ananas*) to be
197 approximately 70 million years old.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

198 **DISCUSSION**

199 The Typhaceae family (order Poales) is diverse, comprising over 50 recognized species
200 (Christenhusz & Byng, 2016). Species in this family are essential components of marshes
201 and wetlands around the world (reviewed in Bansal et al., 2019), and play a role in both
202 bioremediation and the production of biofuel material (Grosshans, 2014; Svedarsky et
203 al., 2016; Bonanno and Cirelli, 2017). This is the first whole genome sequence
204 assembled from a member of the Typhaceae family and adds to the previously
205 characterized chloroplast genome sequences of *Typha latifolia, T. orientalis,* and
206 *Sparganium stoloniferum* (Guisinger et al., 2010; Su et al., 2019; Liu et al., 2020). The
207 assembly and annotation of the Cattail genome, which we note was a key
208 recommendation from a diverse team of *Typha* researchers (Bansal et al., 2019), will be
209 an important tool in wetland management.

210 We explored a variety of short read assemblers given the varying quality of
211 outputs seen in Assemblathon metrics (Bradnam et al., 2013). Of the three SR
212 assemblers used, ABySS produced longer, fewer contigs and scaffolds, which is
213 consistent with previous comparisons (Bradnam et al., 2013). ABySS' de Bruijn
214 approach did not have issues with the high coverage Illumina data (e.g. Richards &
215 Murali, 2015), which we combined with raw long read data using DGB2OLC (Ye et al.
216 2016) as previous plant genomes with similar data showed promising assembly statistics
217 (i.e. Hatakeyama et al., 2018, Zhou et al., 2019, Daccord et al., 2017). The long reads in
218 the DBG2OLC assembly captured more lengthy and repetitive regions, resulting in a
219 sharp rise in contig N50 and contigs 50 Kb or longer (Table 1). However, the backbone
220 of the final assembly was really the PacBio Canu assembly (Tables 1 and S4), with the

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

benefit of this strategy being the removal of chimeras with polished assemblies having very low error rates (Ye et al., 2016).

Accurate genomes are important as sequencing errors can affect both nucleotide diversity and the discovery of markers (Clark & Whittam, 1992). Long-read sequencing further facilitates detection of structural variants (Amarasinghe et al., 2020), which might be useful in studying *Typha* dynamics given the link to hybridization and speciation (Weissensteiner et al., 2020). The final assembly consists of only 1,189 scaffolds, with 95.55% being 50 Kb or longer; *T. latifolia* possessed higher N50 and fewer scaffolds compared to other plant genome assemblies that used PacBio sequencing and similar hybrid assembly pipelines (Reuscher et al., 2018, Redwan et al., 2016, Hatakeyama et al., 2018). With 96.03% of BUSCOs detected and 92.28% of them complete in our assembly, this would suggest a high-quality annotation. The topology of the phylogenomic tree that we reconstructed based on whole genome sequences of six species agrees with previously inferred evolutionary relationships (e.g. Darshetkar et al. 2019) by grouping together the two bromeliad species (*T. latifolia* and *A. comosus*) and the three graminid species (*O. sativa, S. bicolor* and *B. distachyon*), and by inferring a more recent divergence date for graminids compared to bromeliads. Our estimated *Typha* lineage age of approximately 70 million years exceeds an earlier estimate of between 22.64 and 57.60 mya that was based on seven cpDNA markers (Zhou et al., 2018), although is comparable to the estimate of 69.5 Myr that was based on a combination of fossil records and cpDNA sequences (Bremer, 2000). We acknowledge that the use of whole genome sequences to infer divergence times is still in its infancy,

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

and caution that factors such as rate heterogeneity may complicate such inferences (reviewed in Smith et al. 2018).

The *T. latifolia* reference genome now allows for studying *Typha* and hybridization at the molecular level, for example by providing insight into the adaptation of *T. latifolia* populations that may be threatened by climate change, and facilitating research into potentially important processes such as hybrid breakdown in invasive *Typha* hybrids. Additionally, genome-wide markers could provide insight into why the hybrid *T.* x *glauca* is dominant in some areas (Kirk et al., 2011; Freeland et al., 2013) but uncommon in others (Freeland et al., 2013; Ciotir et al., 2017). For example, *T. angustifolia* in North America is thought to have arrived from Europe several centuries ago (Ciotir et al., 2013), and ancestry assessments could test the hypothesis that historical interspecific hybridization led to genetic introgression into some *T. angustifolia* populations, which may help to explain regional erosion of species barriers: locus-specific ancestry estimation analyses, markers, and the identification of potentially adaptive genes can all be used to find evidence of ancient hybridization in *T. angustifolia* (Goulet et al., 2017; Taylor & Larson, 2019). This high-quality draft genome and its comparisons with Poales species will be an indispensable resource for ongoing research into *Typha,* a genus that both sustains and threatens wetlands around the world.

Data Availability Statement: Raw sequencing data can be found in the Short Read Archive (SRA) under accession number PRJNA751759. Genome assembly has been submitted to GenBank (Temporary ID SUB10173221). Code used to generate the data

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

266 can be found at https://gitlab.com/WiDGeT_TrentU/undergrad-theses/-
267 /tree/master/Widanagama_2021. The *T. angustifolia* RNSseq data is in the SRA under
268 SRR15541138. The *T. latifolia* transcriptome used can be found at
269 https://figshare.com/articles/dataset/Typha_latifolia_leaf_transcriptome/5661727/1.

270

271 **Acknowledgements:** We thank Vikram Bhargav for sampling and preparing the *T.*
272 *latifolia* leaves, and Eric Wootton for extracting DNA from the sample. The CanSeq 150
273 project funded the PacBio sequencing: NSERC Discovery grants to JF and AS and two
274 NSERC Undergraduate Student Research Awards to SW supported this research. XX for
275 the *T. angustifolia* transcriptome used in the genome's annotation. And Compute
276 Canada Resources for Research Groups award to AS and Compute Canada support staff
277 supported the bioinformatics.

278

279

280

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

REFERENCES

Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020). Opportunities and challenges in long-read sequencing data analysis. *Genome biology*, *21*(1), 1-16.

Andrews, N. J., & Pratt, D. C. (1978). Energy potential of cattails (*Typha* spp.) and productivity in managed stands. *J. Minn. Acad. Sci.;(United States)*, *44*(2).

Bansal, S., Lishawa, S. C., Newman, S., Tangen, B. A., Wilcox, D., Albert, D., ... & Windham-Myers, L. (2019). *Typha* (Cattail) invasion in North American wetlands: Biology, regional problems, impacts, ecosystem services, and management. *Wetlands*, *39*(4), 645-684.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114-2120.

Bonanno, G., & Cirelli, G. L. (2017). Comparative analysis of element concentrations and translocation in three wetland congener plants: *Typha domingensis, Typha latifolia and Typha angustifolia*. *Ecotoxicology and Environmental Safety*, *143*, 92-101.

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., ... & Korf, I. F. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, *2*(1), 2047-217X.

Bremer, K. (2000). Early Cretaceous lineages of monocot flowering plants. *Proceedings of the National Academy of Sciences*, *97*(9), 4707-4711.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2: Automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR genomics and bioinformatics*, *3*(1), lqaa108.

Bunbury-Blanchette, A., Freeland, J.R., & Dorken, M. (2015). Hybrid *Typha* × *glauca* outperforms native *T. latifolia* under contrasting water depths in a common garden. *Basic and Applied Ecology* 16:394-402.

Chakraborty, M., Baldwin-Brown, J. G., Long, A. D., & Emerson, J. J. (2016). Contiguous and accurate *de novo* assembly of metazoan genomes with modest long read coverage. *Nucleic acids research*, *44*(19), e147-e147.

Christenhusz, M. J., & Byng, J. W. (2016). The number of known plants species in the world and its annual increase. *Phytotaxa*, *261*(3), 201-217.

Chikhi, R., & Medvedev, P. (2014). Informed and automated k-mer size selection for genome assembly. *Bioinformatics*, *30*(1), 31-37.

Ciotir, C., & Freeland, J. (2016). Cryptic intercontinental dispersal, commercial retailers, and the genetic diversity of native and non-native cattails (*Typha* spp.) in North America. *Hydrobiologia*, *768*(1), 137-150.

Ciotir, C., Kirk, H., Row, J. R., & Freeland, J. R. (2013). Intercontinental dispersal of *Typha angustifolia* and *T. latifolia* between Europe and North America has implications for *Typha* invasions. *Biological Invasions*, *15*(6), 1377-1390.

Ciotir, C., Szabo, J., & Freeland, J.R. (2017). Genetic characterization of cattail species and hybrids (*Typha* spp.) in Europe. *Aquatic Botany* 141:51-59.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Clark, A. G., & Whittam, T. S. (1992). Sequencing errors and molecular evolutionary analysis. *Molecular Biology and Evolution*, *9*(4), 744-752.

Daccord, N., Celton, J. M., Linsmith, G., Becker, C., Choisne, N., Schijlen, E., ... & Bucher, E. (2017). High-quality *de novo* assembly of the apple genome and methylome dynamics of early fruit development. *Nature genetics*, *49*(7), 1099-1106.

Darshetkar, A. M., Datar, M. N., Tamhankar, S., Li, P., & Choudhary, R. K. (2019). Understanding evolution in Poales: Insights from Eriocaulaceae plastome. *PLoS One*, *14*(8), e0221423.

Farrer, E. C., & Goldberg, D. E. (2009). Litter drives ecosystem and plant community changes in cattail invasion. *Ecological Applications*, *19*(2), 398-412.

Freeland, J., Ciotir, C., & Kirk, H. (2013). Regional differences in the abundance of native, introduced, and hybrid *Typha* spp. in northeastern North America influence wetland invasions. *Biological Invasions*, *15*(12), 2651-2665.

Geddes, P., Grancharova, T., Kelly, J.J., Treering, D., & Tuchman, N.C. (2014). Effects of invasive *Typha* × *glauca* on wetland nutrient pools, denitrification, and bacterial communities are influenced by time since invasion. *Aquatic Ecology* 48(3):247-258

Goulet, B. E., Roda, F., & Hopkins, R. (2017). Hybridization in plants: old ideas, new techniques. *Plant physiology*, *173*(1), 65-78.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

344 Grace, J. B., & Harrison, J. S. (1986). The biology of Canadian weeds.: 73. *Typha*
345 *latifolia* L., *Typha angustifolia* L. and *Typha* x *glauca* Godr. *Canadian Journal*
346 *of Plant Science*, *66*(2), 361-379.

347 Grosshans, R. (2014). Cattail (*Typha* spp.) biomass harvesting for nutrient capture and
348 sustainable bioenergy for integrated watershed management.

349 Guisinger, M. M., Chumley, T. W., Kuehl, J. V., Boore, J. L., & Jansen, R. K. (2010).
350 Implications of the plastid genome sequence of *Typha* (Typhaceae, Poales) for
351 understanding genome evolution in Poaceae. *Journal of Molecular Evolution*,
352 *70*(2), 149-166.

353 Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith Jr, R. K., Hannick, L. I.,
354 ... & White, O. (2003). Improving the *Arabidopsis* genome annotation using
355 maximal transcript alignment assemblies. *Nucleic acids research*, *31*(19), 5654-
356 5666.

357 Hatakeyama, M., Aluri, S., Balachadran, M. T., Sivarajan, S. R., Patrignani, A., Grüter,
358 S., ... & Shimizu, K. K. (2018). Multiple hybrid *de novo* genome assembly of
359 finger millet, an orphan allotetraploid crop. *DNA Research*, *25*(1), 39-47.

360 Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., ... &
361 Flicek, P. (2021). Ensembl 2021. *Nucleic acids research*, *49*(D1), D884-D891.

362 Humann, J. L., Lee, T., Ficklin, S., & Main, D. (2019). Structural and functional
363 annotation of eukaryotic genomes with GenSAS. In *Gene Prediction* (pp. 29-51).
364 Humana, New York, NY.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

365 Jackman, S. D., Vandervalk, B. P., Mohamadi, H., Chu, J., Yeo, S., Hammond, S. A., ... &
366 Birol, I. (2017). ABySS 2.0: resource-efficient assembly of large genomes using a
367 Bloom filter. *Genome research*, *27*(5), 768-777.

368 Kajitani, R., Toshimoto, K., Noguchi, H., Toyoda, A., Ogura, Y., Okuno, M., ... & Itoh, T.
369 (2014). Efficient *de novo* assembly of highly heterozygous genomes from whole-
370 genome shotgun short reads. *Genome research*, *24*(8), 1384-1395.

371 Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software
372 version 7: improvements in performance and usability. *Molecular biology and*
373 *evolution*, *30*(4), 772-780.

374 Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R.,
375 Ouyang, S., ... & Matsumoto, T. (2013). Improvement of the *Oryza sativa*
376 Nipponbare reference genome using next generation sequence and optical map
377 data. *Rice*, *6*(1), 1-10.

378 Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based
379 genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature*
380 *biotechnology*, *37*(8), 907-915.

381 Kirk, H., Connolly, C., & Freeland, J. R. (2011). Molecular genetic data reveal
382 hybridization between *Typha angustifolia* and *Typha latifolia* across a broad
383 spatial scale in eastern North America. *Aquatic Botany*, *95*(3), 189-193.

384 Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M.
385 (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer
386 weighting and repeat separation. *Genome research*, *27*(5), 722-736.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

387 Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: a resource for
388 timelines, timetrees, and divergence times. *Molecular biology and evolution*,
389 *34*(7), 1812-1819.

390 Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., &
391 Salzberg, S. L. (2004). Versatile and open software for comparing large genomes.
392 *Genome biology*, *5*(2), 1-9.

393 Larkin, D.J., Freyman, M.J., Lishawa, S.C., Geddes, P., Tuchman, N.C., 2012.
394 Mechanisms of dominance by the invasive hybrid cattail *Typha* × *glauca*.
395 *Biological Invasions* 14, 65–77.

396 Lawrence, B. A., Bourke, K., Lishawa, S. C., & Tuchman, N. C. (2016). *Typha* invasion
397 associated with reduced aquatic macroinvertebrate abundance in northern Lake
398 Huron coastal wetlands. *Journal of Great Lakes Research*, *42*(6), 1412-1419.

399 Lawrence, B.A., Lishawa, S.C., Hurst, N., Castillo, B.T., Tuchman, N.C. (2017). Wetland
400 invasion by *Typha* × *glauca* increases soil methane emissions. *Aquatic Botany*
401 137, 80-87

402 Lishawa, S. C., Jankowski, K., Geddes, P., Larkin, D. J., Monks, A. M., & Tuchman, N. C.
403 (2014). Denitrification in a Laurentian Great Lakes coastal wetland invaded by
404 hybrid cattail (*Typha* × *glauca*). *Aquatic sciences*, *76*(4), 483-495.

405 Liu, Z. D., Zhou, X. L., Ma, H. Y., Tian, Y. Q., & Shen, S. K. (2020). Characterization of
406 the complete chloroplast genome sequence of wetland macrophyte *Typha*
407 *orientalis* (Typhaceae). *Mitochondrial DNA Part B*, *5*(1), 136-137.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Lomsadze, A., Burns, P. D., & Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic acids research*, *42*(15), e119-e119.

Lowe, T. M., & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic acids research*, *25*(5), 955–964. https://doi.org/10.1093/nar/25.5.955

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., ... & Wang, J. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience*, *1*(1), 2047-217X.

Miller, R. L., & Fujii, R. (2010). Plant community, primary productivity, and environmental conditions following wetland re-establishment in the Sacramento-San Joaquin Delta, California. *Wetlands Ecology and Management*, *18*(1), 1-16.

Ming, R., VanBuren, R., Wai, C. M., Tang, H., Schatz, M. C., Bowers, J. E., ... & Yu, Q. (2015). The pineapple genome and the evolution of CAM photosynthesis. *Nature genetics*, *47*(12), 1435-1442.

Moscou, M.. (2017). *Typha latifolia leaf transcriptome* (Version 1). figshare. https://doi.org/10.6084/m9.figshare.5661727.v1

Nawrocki, E. P., & Eddy, S. R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*, *29*(22), 2933-2935.

Pieper, S., Dorken, M., & Freeland, J. (2020). Genetic structure in hybrids and progenitors provides insight into processes underlying an invasive cattail (*Typha* × *glauca*) hybrid zone. *Heredity*, *124*(6), 714-725.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Pieper, S. J., Nicholls, A. A., Freeland, J. R., & Dorken, M. E. (2017). Asymmetric hybridization in cattails (*Typha* spp.) and its implications for the evolutionary maintenance of native *Typha latifolia*. *Journal of Heredity*, *108*(5), 479-487.

Redwan, R. M., Saidin, A., & Kumar, S. V. (2016). The draft genome of MD-2 pineapple using hybrid error correction of long reads. *DNA Research*, *23*(5), 427-439.

Reuscher, S., Furuta, T., Bessho-Uehara, K., Cosi, M., Jena, K. K., Toyoda, A., ... & Ashikari, M. (2018). Assembling the genome of the African wild rice *Oryza longistaminata* by exploiting synteny in closely related *Oryza* species. *Communications biology*, *1*(1), 1-10.

Richards, S., & Murali, S. C. (2015). Best Practices in Insect Genome Sequencing: What Works and What Doesn't. *Current opinion in insect science*, *7*, 1–7. https://doi.org/10.1016/j.cois.2015.02.013

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210-3212.

Shen, W., Le, S., Li, Y., & Hu, F. (2016). SeqKit: a cross-platform and ultrafast toolkit for FASTA/Q file manipulation. *PloS one*, *11*(10), e0163962.

Smit, AFA, Hubley, R. (2008-2015). *RepeatModeler* (Open-1.0). Institute for Systems Biology. http://www.repeatmasker.org

Smit, AFA, Hubley, R & Green, P. (2013-2015). *RepeatMasker* (Open-4.0). Institute for Systems Biology. http://www.repeatmasker.org

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Smith, S. A., Brown, J. W., & Walker, J. F. (2018). So many genes, so little time: A practical approach to divergence-time estimation in the genomic era. *PloS one*, *13*(5), e0197433.

Smith, S. G. (1987). *Typha*: its taxonomy and the ecological significance of hybrids. *Archiv für Hydrobiologie*, *27*(1), 129-138.

Snow, A. A., Travis, S. E., Wildová, R., Fér, T., Sweeney, P. M., Marburger, J. E., ... & Mutegi, E. (2010). Species- specific SSR alleles for studies of hybrid cattails (*Typha latifolia* × *T. angustifolia*; Typhaceae) in North America. *American Journal of Botany*, *97*(12), 2061-2067.

Solares, E. A., Chakraborty, M., Miller, D. E., Kalsow, S., Hall, K., Perera, A. G., ... & Hawley, R. S. (2018). Rapid low-cost assembly of the *Drosophila melanogaster* reference genome using low-coverage, long-read sequencing. *G3: Genes, Genomes, Genetics*, *8*(10), 3143-3154.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312-1313.

Su, T., Yang, J. X., Lin, Y. G., Kang, N., & Hu, G. X. (2019). Characterization of the complete chloroplast genome of *Sparganium stoloniferum* (Poales: Typhaceae) and phylogenetic analysis. *Mitochondrial DNA Part B*, *4*(1), 1402-1403.

Svedarsky, D., Bruggman, J., Ellis-Felege, S., Grosshans, R., Lane, V., Norrgard, R., ... & Brenny, T. (2016). Cattail management in the Northern Great Plains: Implications for wetland wildlife and bioenergy harvest. *St. Paul*.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

472 Tangen, B. A., Bansal, S., Freeland, J. R., Travis, S. E., Wasko, J. D., McGonigle, T. P., ...
473 & Meier, J. A. (2021). Distributions of native and invasive *Typha* (cattail)
474 throughout the Prairie Pothole Region of North America. *Wetlands Ecology and*
475 *Management*, 1-17.

476 Taylor, S. A., & Larson, E. L. (2019). Insights from genomes into the evolutionary
477 importance and prevalence of hybridization in nature. *Nature ecology &*
478 *evolution*, *3*(2), 170-177.

479 Tisshaw, K., Freeland, J., & Dorken, M. (2020). Salinity, not genetic incompatibilities,
480 limits the establishment of the invasive hybrid cattail *Typha* × *glauca* in coastal
481 wetlands. *Ecology and evolution*, *10*(21), 12091-12103.

482 Tuchman, N. C., Larkin, D. J., Geddes, P., Wildova, R., Jankowski, K., & Goldberg, D. E.
483 (2009). Patterns of environmental change associated with *Typha* x *glauca*
484 invasion in a Great Lakes coastal wetland. *Wetlands*, *29*(3), 964-975.

485 Wang, Y., Coleman-Derr, D., Chen, G., & Gu, Y. Q. (2015). OrthoVenn: a web server for
486 genome wide comparison and annotation of orthologous clusters across multiple
487 species. *Nucleic acids research*, *43*(W1), W78-W84.

488 Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., ... & Earl, A.
489 M. (2014). Pilon: an integrated tool for comprehensive microbial variant
490 detection and genome assembly improvement. *PloS one*, *9*(11), e112963.

491 Waterhouse, R. M., Tegenfeldt, F., Li, J., Zdobnov, E. M., & Kriventseva, E. V. (2013).
492 OrthoDB: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic*
493 *acids research*, *41*(D1), D358-D365.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

494 Weissensteiner, M. H., Bunikis, I., Catalán, A., Francoijs, K. J., Knief, U., Heim, W., ... &
495 Wolf, J. B. (2020). Discovery and population genomics of structural variation in a
496 songbird genus. *Nature communications*, *11*(1), 1-11.

497 Xu, L., Dong, Z., Fang, L., Luo, Y., Wei, Z., Guo, H., Zhang, G., Gu, Y. Q., Coleman-Derr,
498 D., Xia, Q., & Wang, Y. (2019). OrthoVenn2: a web server for whole-genome
499 comparison and annotation of orthologous clusters across multiple species.
500 *Nucleic acids research*, *47*(W1), W52–W58. https://doi.org/10.1093/nar/gkz333

501 Xu, Z., Feng, Z., Yang, J., Zheng, J., & Zhang, F. (2013). Nowhere to invade: *Rumex*
502 *crispus* and *Typha latifolia* projected to disappear under future climate
503 scenarios. *PLoS One*, *8*(7), e70728.

504 Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum
505 likelihood. *Computer applications in the biosciences*, *13*(5), 555-556.

506 Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular*
507 *biology and evolution*, *24*(8), 1586-1591.

508 Ye, C., Hill, C. M., Wu, S., Ruan, J., & Ma, Z. S. (2016). DBG2OLC: efficient assembly of
509 large genomes using long erroneous reads of the third generation sequencing
510 technologies. *Scientific reports*, *6*, 31900.

511 Zedler, J.B. & Kercher, S. (2004). Causes and consequences of invasive plants in
512 wetlands: Opportunities, opportunists, and outcomes. *Critical Reviews in Plant*
513 *Sciences,* 23 (5), 431-452

514 Zhou, B., Yu, D., Ding, Z., & Xu, X. (2016). Comparison of genetic diversity in four
515 *Typha* species (Poales, Typhaceae) from China. *Hydrobiologia*, *770*(1), 117-128.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

516 Zhou, B., Tu, T., Kong, F., Wen, J., & Xu, X. (2018). Revised phylogeny and historical
517 biogeography of the cosmopolitan aquatic plant genus *Typha* (Typhaceae).
518 *Scientific reports*, *8*(1), 1-7.

519 Zhou, Y., Minio, A., Massonnet, M., Solares, E., Lv, Y., Beridze, T., ... & Gaut, B. S.
520 (2019). The population genetics of structural variants in grapevine domestication.
521 *Nature plants*, *5*(9), 965-979.

522

523

524

525

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

TABLES

Table 1. Genome assembly statistics of our four-step hybrid assembly. Short read (SR) contig assembly consists of contigs assembled from 100% of the Illumina reads using the ABySS assembler. The DBG2OLC assembler combined the 100% ABySS contigs and PacBio long reads. The long read (LR) contig assembly was generated from only PacBio long reads using Canu. The merged polished scaffolds are from merging the DBG2OLC assembly and LR contigs and were polished using Pilon.

	SR contig assembly – step 1	DB2OLC contig assembly – step 2	LR contig assembly – step 3	Merged polished scaffolds – step 4
Genome Size (Mb)	263.74	193.16	287.63	287.80
Contigs/ Scaffolds	365,565	1,840	1,190	1,189
N50 / L50	11.43 Kb / 5,314	132.07 Kb / 412	8.71 Mb / 13	8.71 Mb / 13
N90 / L50	201 bp / 193,744	52.95 Kb / 1,358	58.92 Kb / 530	58.94 Kb / 529
Max sequence Length	154.73 Kb	934.40 Kb	18.70 Mb	18.69 Mb
Scaffolds > 10 Kb	6,048	1,833	1,140	1,139
Scaffolds > 25 Kb	1,759	1,785	1,127	1,126
Scaffolds > 50 Kb	362	1,445	821	821

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

% of scaffolds > 50 Kb	9.46	92.34	95.54	95.55
GC content (%)	38.40	38.50	38.11	38.11

533

534

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

535 **Table 2.** Summary of repeats masked in the *Typha latifolia* genome.

	Length (bp)	Percentage of genome (%)
SINE	0	0
LINE	3,499,426	1.22
LTR elements	44,172,983	15.35
DNA elements	3,735,882	1.30
Unclassified	68,736,281	23.88
Small RNA	0	0
Satellites	0	0
Simple repeats	5,689,134	1.98
Low complexity	670,871	0.23
Total	126,178,695	43.84

536

537

538

539

540

541

542

543

544

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

FIGURE CAPTIONS

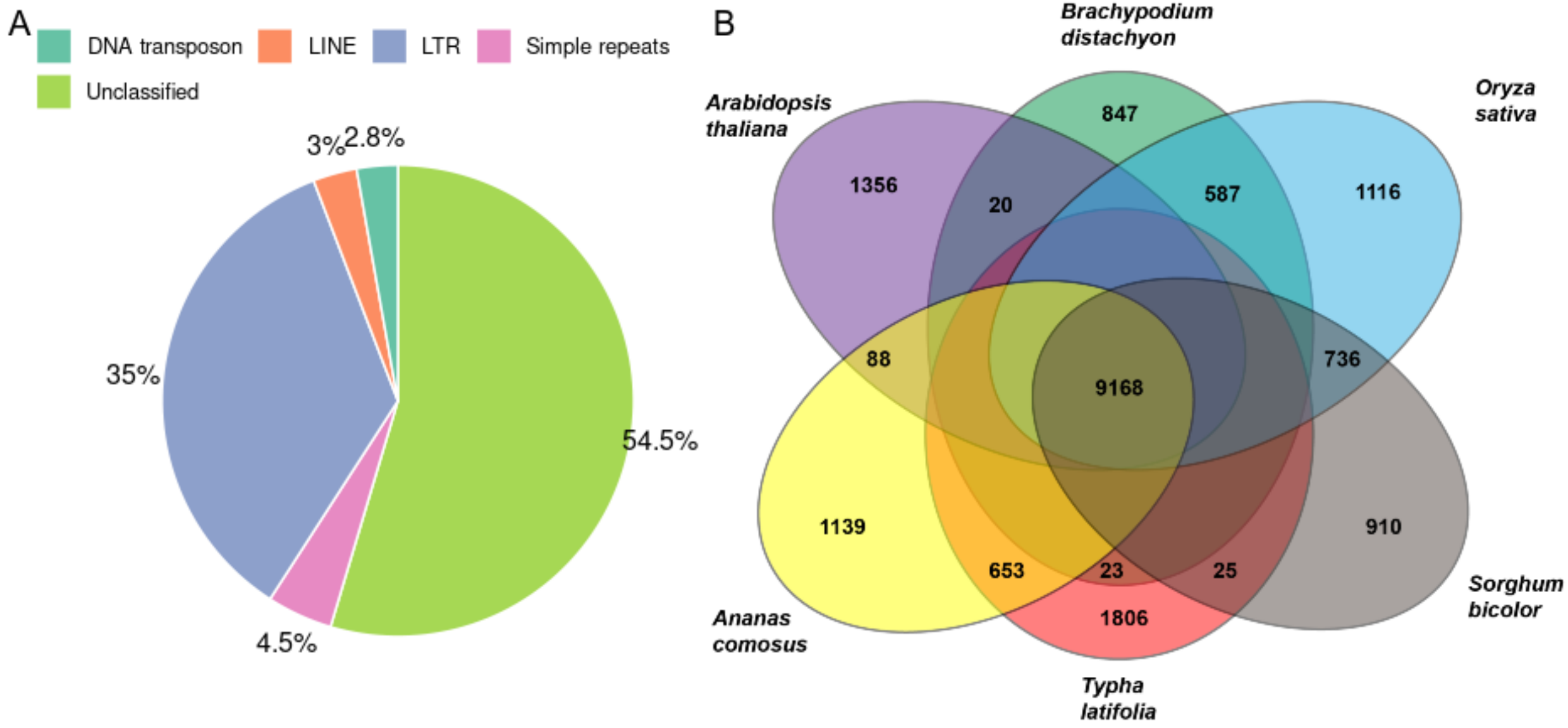
Figure 1. Broadleaf cattail (*Typha latifolia*). Photo by Joanna Freeland.

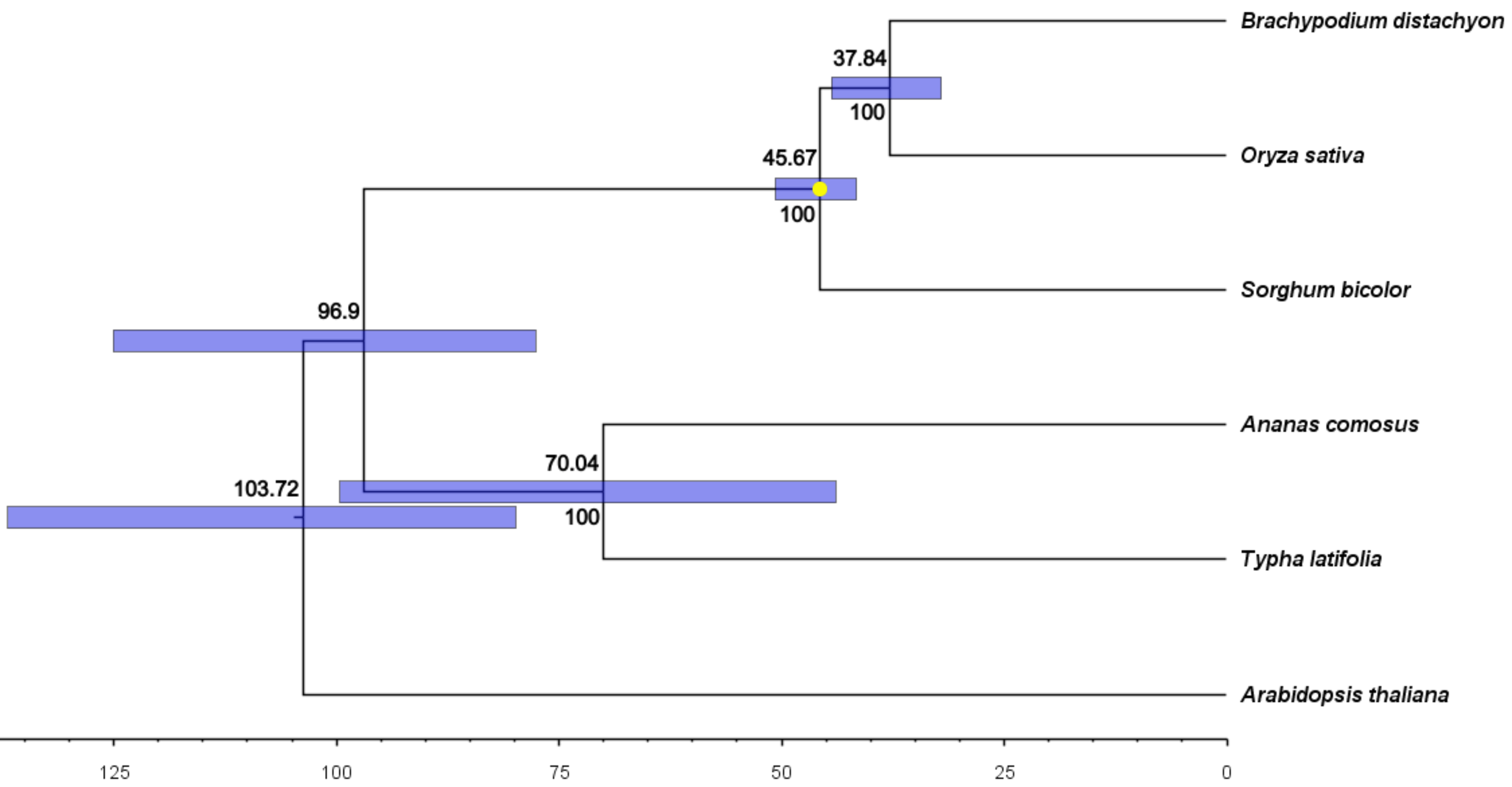
Figure 2. a) Percentages of repeat types masked in the *T. latifolia* genome. Types of repeats include DNA transposons, long interspersed nuclear elements (LINEs), long terminal repeats (LTRs), unclassified repeats, and simple repeats. b) Venn diagram of orthologous gene clusters among the broadleaf cattail (*Typha latifolia*), pineapple (*Ananas comosus*), thale cress (*Arabidopsis thaliana*), stiff brome (*Brachypodium distachyon*), rice (*Oryza sativa Japonica*), and broom-corn (*Sorghum bicolor*). Only the numbers of ortholog clusters of adjacent species, those common to all species, and those unique to each species are labelled.

Figure 3. Phylogenetic tree with divergence times, based on the alignment of 1,900 single-copy gene clusters. 95% credible divergence times are shown as blue bars and were estimated using MCMCTree. Divergence times and bootstrap values are shown above and below the nodes, respectively. The yellow dot indicates the calibration point.

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.23.457420; this version posted August 24, 2021
(which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to disp
available under a CC-BY 4.0 International license




A
DNA transposon
LINE
LTR
Simple repeats
Unclassified
3%
2.8%
35%
54.5%
4.5%
B
Brachypodium distachyon
Arabidopsis thaliana
Oryza sativa
Sorghum bicolor
Ananas comosus
Typha latifolia
1356
847
20
587
1116
88
9168
736
1139
910
653
23
25
1806


Brachypodium distachyon
37.84
100
Oryza sativa
45.67
100
Sorghum bicolor
96.9
Ananas comosus
70.04
103.72
100
Typha latifolia
Arabidopsis thaliana
125
100
75
50
25
0