Welcome to the big leaves: best practices for improving genome annotation in non-model plant genomes

Vidya S Vuruputoor[1], Daniel Monyak[1], Karl C. Fetter[1], Cynthia Webster[1], Akriti Bhattarai[1], Bikash Shrestha[1], Sumaira Zaman[1], Jeremy Bennett[1], Susan L. McEvoy[1], Madison Caballero[1] Jill L. Wegrzyn[1]

[1] Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269

Author for correspondence: jill.wegrzyn@uconn.edu

ABSTRACT

- **Premise of the study**: Robust standards to evaluate quality and completeness are lacking for eukaryotic structural genome annotation. Genome annotation software is developed with model organisms and does not typically include benchmarking to comprehensively evaluate the quality and accuracy of the final predictions. Plant genomes are particularly challenging with their large genome sizes, abundant transposable elements (TEs), and variable ploidies. This study investigates the impact of genome quality, complexity, sequence read input, and approach on protein-coding gene prediction.
- **Methods:** The impact of repeat masking, long-read, and short-read inputs, *de novo*, and genome-guided protein evidence was examined in the context of the popular BRAKER and MAKER workflows for five plant genomes. Annotations were benchmarked for structural traits and sequence similarity.
- **Results:** Benchmarks that reflect gene structures, reciprocal similarity search alignments, and mono-exonic/multi-exonic gene counts provide a more complete view of annotation accuracy. Transcripts derived from RNA-read alignments alone are not sufficient for genome annotation. Gene prediction workflows that combine evidence-based and *ab initio* approaches are recommended, and a combination of short and long-reads can improve genome annotation. Adding protein evidence from *de novo assemblies*, genome-guided transcriptome assemblies, or full-length proteins from OrthoDB generates more putative false positives as implemented in the current workflows. Post-processing with functional and structural filters is highly recommended.
- **Discussion:** While annotation of non-model plant genomes remains complex, this study provides recommendations for inputs and methodological approaches. We discuss a set

of best practices to generate an optimal plant genome annotation, and present a more robust set of metrics to evaluate the resulting predictions.

Keywords: genome annotation, plant genomes, gene identification, BRAKER, TSEBRA, MAKER, StringTie2

INTRODUCTION

The first published plant genome, *Arabidopsis thaliana*, was released in 2000 (Arabidopsis Genome Initiative, 2000). Its small genome size (135Mb) and minimal repeat content stand in stark contrast to the plant species sequenced and assembled today (Kress et al., 2022). NCBI's (https://www.ncbi.nlm.nih.gov/) genome repository contains genomes of over 900 land plant species, and roughly half of these are assembled to chromosome scale. The total number of complete reference plant genomes has more than doubled in the last five years (Marks et al., 2021). Initiatives like the Open Green Genomes (OGG) (https://phytozome-next.jgi.doe.gov/ogg/), 10KP (Cheng et al., 2018), and the Earth BioGenome Project (Lewin et al., 2022) are improving the phylogenetic representation of plant genomes by sampling underrepresented clades. The plant genomes published today are more likely to be polyploids and/or larger genomes with substantial transposable element content (Sun et al., 2022). The combination of high throughput sequencing advancements, particularly long reads and chromosome conformation capture approaches have enabled the completion of these more challenging assemblies (Pucker et al., 2022).

While genome assembly has seen substantial improvements in accuracy and contiguity, structural annotation remains challenging. This process delineates the physical positions of genomic features, including protein-coding genes, promoters, and regulatory elements. It can be followed by functional annotation, which assigns biological descriptors to the identified features. The accurate classification of these features provides the basis for questions focused on species evolution, population dynamics, and functional genomics. Errors in genome annotation are frequent, even among well-studied models, and are propagated through downstream analyses (Deutekom et al., 2019; Meyer et al., 2020; Salzberg, 2019). In most eukaryotes, genome annotation is challenged by partial conservation of sequence patterns, variable lengths

of introns, variable distances between genes, alternative splicing, and higher densities of TEs and pseudogenes (Kersey, 2019; Salzberg, 2019). As a result of these complexities, the structural annotation process requires more advanced informatic tools and skills that support the integration and manipulation of large datasets (Mudge & Harrow, 2016).

Structural and functional genome annotation proceeds in three stages: identifying and masking noncoding regions (repeats); predicting physical positions of gene structures; and assigning biological information to the predictions (Jung et al., 2020). Repeat regions are soft-masked (eg., RepeatMasker (Smit, AFA, Hubley, R & Green, P., 2013-2015) and RepeatModeler2 (Flynn et al., 2020)), which means these regions are indicated but not obscured to annotation software. This is followed by gene prediction, which may be *ab initio* (evidence-free) or evidenced-based. Evidence-based approaches use RNA-Seq and protein sequence similarity search alignments. Evidence-based approaches are often used in combination with *ab initio* (e.g. AUGUSTUS; (Stanke & Waack, 2003)) to generate models that are trained on patterns associated with true genes. Given the advanced state of high throughput transcriptome sequencing, it is common to resolve transcripts from RNA reads through genome-guided approaches, such as StringTie2 (Kovaka et al., 2019). Long-read cDNA sequencing through PacBio and Oxford Nanopore can provide additional resolution and improve the identification of splice variants. When extrinsic evidence from RNA-seq and protein alignments are available, workflow packages like MAKER (Campbell, Holt, et al., 2014; Cantarel et al., 2008; Holt & Yandell, 2011) and BRAKER (Brůna et al., 2021; Hoff et al., 2016, 2019) can assist in training *ab initio* prediction tools. These packages can leverage sequence data from the target species as well as external evidence from closely related species. While these workflows can simplify the integration across external evidence, downstream packages are still required to select or modify the resulting predictions (Banerjee et al., 2021; Gabriel et al., 2021; Haas et al., 2008).

52 Here, we provide a comprehensive evaluation of plant genome annotation workflows,
53 intentionally selecting beyond the typical model species to represent some of the more complex
54 genomes under investigation today. In doing so, we evaluate the impact of repeat-masking
55 using two different implementations of the RepeatModeler2 framework (Flynn et al., 2020). This
56 is followed by exploring the role of read length and accuracy, and the impact of short-read and
57 long-read data. Finally, we examine the contribution of protein evidence, generated from d*e*
58 *novo* assembly of the RNA inputs and a genome-guided assembly. These variations are
59 examined in the MAKER and BRAKER frameworks to emphasize the importance of defining
60 benchmarks to guide downstream filtering approaches. Finally, the largest and most repetitive
61 genome assessed in this study, *Liriodendron chinense*, was used to demonstrate best practices
62 to refine the predictions.

63 METHODS

64

65 ***Gathering plant genome datasets-***

Five plant genomes were chosen for this study, including Chinese tuliptree (*Liriodendron chinense*) (Chen et al. 2019), black cottonwood (*Populus trichocarpa* v3) (Tuskan et al. 2006), Chinese rose (*Rosa chinensis*) (Raymond et al. 2018), thale cress (*Arabidopsis thaliana* TAIR 10) (Cheng et al. 2017), and a bryophyte, the common cord-moss (*Funaria hygrometrica*) (Kirbis et al., 2022) (Table S1). The genomes were selected to represent two model systems (*Populus* and *Arabidopsis*) with well curated structural annotations and three non-model systems that exclusively used computational techniques to produce the annotations. Two of these non-models were also more divergent examples, representing the only sequenced member of their genus (*Funaria* and *Liriodendron*). The public assembly and annotation for each species were accessed from NCBI and genome completeness was estimated by searching the genome and annotation for the conserved single-copy orthologs in the Embryophyta odb10 BUSCO v.5.0.0 (Simão et al., 2015). The contiguity of the reference genomes was assessed with Quast v5.0.2 (Gurevich et al., 2013). Published annotation files were summarized with gFACs (Caballero and Wegrzyn, 2019).

Read sets available through NCBI's Sequence Read Archive (SRA) were accessed to provide transcriptomic evidence for each species and included a variety of tissue types. The Illumina short-read libraries were sequenced with Illumina HiSeq 2500 (100bp paired-end). The read sets included at least four libraries, between 20-82M reads before quality control (QC), and a minimum of 16M reads after QC. Pacific Biosciences Iso-Seq long-reads were accessed for *Populus* and *Liriodendron*, and Oxford PromethION reads were available for *Rosa* and *Arabidopsis*. The read sets for long-read data ranged between 161K-41M total reads per species (Table S2).

Repeat masking and read alignment-

91 RepeatModeler2 (Flynn et al., 2020) was used to construct repeat libraries with default settings,
92 and repeats were soft masked with the libraries constructed via RepeatMasker v.4.0.6 (Smit et
93 al. 2013-2015). The genomes of *Arabidopsis, Funaria, Populus, and Liriodendron* were
94 additionally masked using RepeatModeler2 with additional LTR identification (-LTRStruct flag).
95 Quality assessment of the Illumina short reads was performed using FastQC v.0.11.7 (Andrews,
96 2010) before and after trimming low-quality bases. Sickle v.1.33 (Joshi NA, 2011) was used to
97 trim low-quality bases with 50bp as the minimum read length threshold. Single end reads
98 generated post trimming were excluded from RNA alignments and assembly. The trimmed short
99 reads were aligned against their reference genomes using HISAT2 v2.2.0 (Kim et al., 2019).
100 HISAT2 was selected for its performance in recent benchmarking studies and as the aligner of
101 choice for input to Stringtie2 (Corchete et al., 2020; Musich et al., 2021). Long-read RNA data
102 were obtained for four species: *Arabidopsis* and *Rosa* were sequenced with Oxford Nanopore,
103 and *Populus* and *Liriodendron* were sequenced with PacBio Sequel. The long-read data sets
104 were aligned against their respective genomes using Minimap2 v2.1.7 (Li, 2018, 2021).
105
106 ***Generation of protein evidence-***
107 To generate protein evidence, Illumina short reads were assembled *de novo* using Trinity
108 v.2.8.5 with a minimum contig length of 300 bp (Grabherr et al., 2011). The assembled
109 transcriptomes for the multiple libraries were combined, and putative coding regions were
110 predicted using TransDecoder v.5.3.0 (http://transdecoder.github.io). TransDecoder is one of
111 several frame-selection methods available and performs in a comparable manner but not
112 always superior in all metrics (Bolger et al., 2018). For this study, it was selected as the most
113 widely used package for this purpose. Redundancy in the predicted coding regions was reduced
114 after clustering at 98% identity using UCLUST, a clustering algorithm of USEARCH v.9.0.2132
115 (Edgar, 2010). Frame-selected transcripts shorter than 300 bp were removed. The remaining
116 transcripts were aligned to the genome using GMAP v.2019-06-10 (Wu & Watanabe, 2005).

The predicted proteins (from the same Transdecoder run) were aligned to the reference genome using GenomeThreader v 1.7.1 (Gremme, 2014).

To provide protein evidence from genome-guided sources, the previously aligned Illumina short-reads (via HISAT2) were constructed into transcripts with StringTie2 v2.2.0 (Kovaka et al., 2019; M. Pertea et al., 2015). Long-reads were treated similarly, along with a combination of short and long-reads. The predicted transcripts were extracted using gffRead (G. Pertea & Pertea, 2020) and frame selected with TransDecoder. The transcriptome alignment annotation file (gff3) was passed to gFACs for evaluation of gene model statistics. Completeness of the aligned transcripts and protein sequences were estimated using BUSCO. Finally, to provide evidence from external sources (not derived from any transcriptomic inputs), full-length protein evidence from OrthoDB (odb10_plants) was provided for BRAKER/TSEBRA runs.

Genome annotations-

Each genome was tested in four primary open-source annotation softwares to predict gene models (Table 1). Several different runs of BRAKER v.2.1.5 (Hoff et al., 2019) and BRAKER/TSEBRA (Gabriel et al., 2021) were used with various combinations of RNA-Seq (long and short-read inputs) and protein evidence. MAKER v.3.1.3 (Cantarel et al., 2008) was run once with transcript and protein evidence. Finally, StringTie2 (Kovaka et al., 2019), with TransDecoder, was used to generate genome-guided predictions from RNA evidence alone (Table S12).

MAKER annotation-

MAKER (MK) was run on the soft-masked reference genomes of *Arabidopsis, Populus,* and *Funaria* with repeats estimated using the additional LTR detection method in RepeatModeler2 (LTRStruct flag; RM2+). This was intended to emulate the MAKER-P (Campbell, Law, et al., 2014) method since the original repeat and pseudogene identification protocols are deprecated. MK (RM2+) was executed (i.e., trained) twice. The annotations derived from MK (RM2+) used protein evidence generated from *de novo* assembled RNA-reads from Trinity. These models were used to train *ab initio* gene prediction software AUGUSTUS v.3.3.3 (Stanke & Waack, 2003) and SNAP v. 2006-07-28 ((Korf, 2004). The Hidden Markov Models (HMMs) trained using AUGUSTUS and SNAP were used along with initial aligned evidence (est2genome and protein2genome parameters) for the second MK (RM2+) run to generate the final gene models.

Assessment of gene predictions-

The quality of genome annotation among different gene prediction methods was evaluated with three primary metrics: (1) the mono-exonic (single-exon) and multi-exonic (multiple exon) ratio; (2) conserved single-copy orthologs queried from the predicted gene models using BUSCO (embryophyta database v10), and (3) gene prediction assessment with EnTAP v0.10.8 (Hart et al., 2020) using a 70% reciprocal functional annotation approach with NCBI's Refseq Plant and Uniprot databases. The mono:multi ratio was calculated from the gFACs summary report run with default parameters (Caballero & Wegrzyn, 2019). We regard a mono: multi ratio near 0.2 to be ideal and have further validated this with a larger set of model plant genomes (Table S3) (Jain et al. 2008). The gene prediction assessment was recorded as a percentage of sequence similarity hits to the total number of genes. The value of the reciprocal BLAST search will be dependent on the phylogenetic relationships of the target species. While higher annotation rates indicate better gene prediction assessment, species-specific genes will always be missed

(Armisén et al., 2008), therefore a minimum annotation rate of 80% is a reasonable threshold to expect when dealing with most non-models. Similarly, a higher BUSCO score indicates a better annotation since BUSCO utilizes OrthoDB to form its conserved sets and the recommended target score is >95% for land plants (Manni et al., 2021; Simão et al., 2015).

The sensitivity and precision of the runs for *Arabidopsis* and *Populus* were assessed using Mikado v2.3.2 (Venturini et al., 2018), by comparing the predicted gene models to the current reference annotations.

Post-processing filtering-

The predicted gene models for *Liriodendron* were taken a step further to refine the genome annotation. Post-process filtering was performed using gFACs and assessed for improvement using BUSCO completeness scores and annotation rates. The mono-exonic and multi-exonic genes predicted for *Liriodendron* were filtered for unique genes. The mono-exonic genes were further filtered for the presence of protein domains using InterProScanv.5.35-74.0 and Pfam (Jones et al., 2014; Quevillon et al., 2005). Multi-exonic genes that did not have an EggNOG or a sequence similarity hit were removed, and the final annotation was assessed using gFACs and EnTAP.

RESULTS

46 ***Genome sizes, repeats, and published annotations-***

47 The genome sizes of the five species assessed represented a 10-fold difference between the
48 smallest genome of *Arabidopsis* (~119 Mb) and the largest of *Liriodendron* (~1.7 Gb) (Fig 1A;
49 Table 2). *Liriodendron* (73.18%) and *Rosa* (60.58%) have higher levels of repeat content, and
50 *Arabidopsis* has the lowest (23.9%). *Arabidopsis* is the most complete chromosome-scale
51 genome, with seven contigs reflecting its five chromosomes and two organellar chromosomes.
52 The other genomes are assembled into pseudochromosomes (except for *Liriodendron*). Once
53 the genomes were downloaded, contigs < 500 bp were removed. The published genome
54 assemblies and annotations were compared in terms of completeness via BUSCO (Fig 1, Table
55 2). When BUSCO is run in genome mode, it searches the genome for the set of 1614 single-
56 copy orthologs in the embryophyte database. Aside from *Funaria*, which had the lowest
57 completeness score of 82.4%, the remaining plant genomes ranged from 94% to 99%. When
58 we evaluated the published annotations for the same species, and ran BUSCO in protein mode,
59 a slight decrease in completeness was observed in every species except *Funaria* and
60 *Arabidopsis* (Fig 1B). The largest reduction in BUSCO score was observed in *Liriodendron*
61 (98.6% to 75.1%). The discrepancy between the estimated completeness at the genome-level
62 and most of the published annotations speaks to the challenges of achieving an accurate
63 structural annotation.
64
65 RepeatModeler2 (RM2) with and without the LTRStruct package (the additional LTR masking
66 module) (Flynn et al. 2020) was used to soft mask repeats in four of the genomes. The increase
67 in repeat content was marginal in all species, ranging from 1% in *Funaria* to 5% in *Populus*.
68 Comparisons using the LTRStruct flag were denoted as RM2+ (Table S4).
69

70 ***Transcriptome evidence-***

71 For the subsequent genome annotation analysis, the Illumina RNA short reads were first aligned
72 to the genome. All libraries, ranging from four to 20 per species, aligned at over 97%, apart from
73 *Rosa* (92%) (Table 3; Table S5). Long-read RNA libraries were aligned with Minimap2 for four
74 species: *Arabidopsis* (Nanopore reads at 97.1%), *Populus* (Iso-Seq reads at 92.01%),
75 *Liriodendron* (Iso-Seq reads at 95.5%), and *Rosa* (Nanopore reads at 99%). The N50s for the
76 long reads range from 976 bp in *Rosa* to 4.6 Kb in *Liriodendron* (Table S6).
77

78 ***Transcript-derived annotations-***

79 The reads were assembled using StringTie2 (ST2) and Trinity. Trinity *de novo* assemblies of the
80 Illumina short reads generated longer transcripts, with N50s ranging from 1.2 Kb (151,265
81 transcripts in total) in *Funaria*, to 3.06 Kb (2,839,867 transcripts) in *Liriodendron*. Among
82 genome-guided assemblies with StringTie2 (ST2(SR)), the range was much smaller, with N50s
83 ranging from 369 bp (59,741 transcripts in total) in *Funaria* to 2.54 Kb (37,747 transcripts) in
84 *Arabidopsis* (Table S6). The StringTie2 (ST2(LR) and ST2(SR/LR)) range was longer, with
85 N50s ranging from 1.07 Kb (20,633 transcripts in total) in *Rosa* ST2(LR) to 2.36 Kb (45,785
86 transcripts) in *Liriodendron* ST2 (SR/LR) (Table 3; Table S6). The StringTie2 and Trinity
87 transcripts were aligned back to the genome using GMAP after frame-selection. BUSCO scores
88 for the aligned transcriptomes derived from short read data, run in transcriptome mode, ranged
89 from 73% in *Funaria* to 83% in *Rosa* for Trinity, and 73% in *Liriodendron* to 97% in *Rosa* using
90 StringTie2 (Table 3). The BUSCO scores were the lowest for the ST2 (LR) runs across all
91 species as compared to the other StringTie2 only runs. For the ST2 (SR/LR), the BUSCO
92 scores were lower than ST2 (SR), except for *Rosa*, where the ST2 (SR/LR) was 97.2% as
93 opposed to 97% in ST2 (SR). In all species, ST2 (SR/LR) had higher BUSCO scores than ST2
94 (LR). Despite Trinity producing more than double the total transcripts than StringTie2, the

95 BUSCO completeness score of most StringTie2 runs were much higher than that of Trinity.
96 *Liriodendron* remained the only exception with a slightly higher BUSCO score from Trinity.
97
98 *Arabidopsis* and *Populus* were further evaluated with Mikado to compare the sensitivity and
99 specificity of published annotations (Fig 4B, Table S7). Overall, StringTie2 predictions had
100 higher sensitivity and precision rates compared to the Trinity runs. From this point, Trinity was
101 excluded, and StringTie2 runs were compared against BRAKER and TSEBRA predictions (Fig
102 4B).
103
104 The mono: multi ratios produced by StringTie2 ranged from 0.15 in *Populus* (ST2 (LR)) to 0.53
105 in *Liriodendron* (ST2 (LR)), which were an improvement over the mono: multi ratios produced
106 from the BRAKER annotations that ranged from 0.37 in *Arabidopsis* (BR (LR)) to 1.27 in *Funaria*
107 (BR (SR/RM2+)). The BUSCO scores of the proteins predicted from BRAKER were generally
108 higher than the BUSCO scores from StringTie2. For example, *Arabidopsis* StringTie2 runs
109 range from 85% (ST2 (LR)) to 95.5% in ST2 (SR), and BRAKER runs ranged from 94% (BR
110 (LR)) to 95.9% (BR (SR)). However, some runs are comparable, the ST2 (SR) run with a
111 BUSCO score of 95% was like the BR (SR) run at 95% and the BR (SR/RM2+) run at 95% in
112 *Arabidopsis*. StringTie2 predicted models had a higher annotation rate, in general, compared to
113 BRAKER. For example, the EnTAP annotation rate in *Funaria* was just over 40% post BRAKER
114 but was near 60% from the StringTie2 runs (Fig 2).
115
116 ***Genome annotation with MAKER -***
117 To replicate MAKER-P's repeat pipeline, the RM2+ genome was used for *Arabidopsis, Populus*,
118 and *Funaria* for the MAKER runs. BUSCO completeness was low, compared with BRAKER
119 runs, and ranged from 19.6% in *Populus* to 90.4% in *Arabidopsis* (Fig 3A). The mono: multi ratio
120 of MAKER(RM2+) for *Arabidopsis* was comparable to the BRAKER runs for the same species

121 (0.22 for BR (SR) and BR (SR/RM2+), 0.24 for BR (LR), and 0.23 for BR (SR/LR)). The MK
122 (RM2+) predictions for the total number of genes in *Arabidopsis* and *Funaria* were in the
123 expected range for these species, 22K and 44K genes, respectively, whereas only 7K genes
124 were predicted for *Populus*. The gene lengths ranged from 1.8 Kb in *Funaria* to 2.3 Kb in
125 *Arabidopsis* (Table S8). The best run for MK (RM2+) was for *Arabidopsis*, with a mono: multi
126 ratio of 0.22 and a BUSCO score of 90.4%. On the other hand, the mono: multi ratio for *Populus*
127 was 0.07, and the BUSCO score was 19.6%.

128 The model systems, *Arabidopsis* and *Populus*, further were evaluated with Mikado to compare
129 the sensitivity and specificity of the published annotations (Fig 3B; Table S7). The sensitivity
130 and precision scores for gene predictions were the lowest from MAKER, followed by Trinity, and
131 were the highest from TSEBRA. StringTie2 and BRAKER yielded similar sensitivity and
132 specificity scores for *Arabidopsis*, whereas for *Populus* the sensitivity score was lower than
133 those from BRAKER runs. Given its overall low scores, MAKER was excluded from the
134 subsequent comparisons. It should be noted, however, that protocols are available that
135 recommend using GeneMark in MAKER. With this addition to the training protocol, MAKER was
136 shown to have a higher accuracy in prediction (Brůna et al., 2021).
137

138 ***Genome annotation with BRAKER: -***

139 In general, BUSCO scores were improved in the BRAKER and TSEBRA runs, with the
140 TSB(SR/OrthoDB) predictions scoring the highest. The best BUSCO scores were from
141 TSB(SR/OrthoDB), ranging from 98.9% (TSB; SR/OrthoDB) in *Arabidopsis* to 82.2% in *Funaria*.
142 Overall, the TSEBRA runs fared better with the annotation rates ranging from 99% in
143 *Arabidopsis* (TSB; SR/OrthoDB) to 53% in *Funaria*. On the other hand, the TSEBRA runs had
144 the worst mono:multi ratios. For example, TSB (SR/ST2/RM2+) has the highest ratio at 1.27 for
145 *Funaria* (Fig 4).

Overall, the gene models generated by BRAKER for *Arabidopsis* performed similarly across runs when evaluated by BUSCO scores. The mono: multi ratios across BRAKER runs ranged between 0.23 to 0.39, and the annotation rates were consistently above 95%.

The annotation rates for *Funaria* were lower than expected, 43% for BR (SR) and 53% for nearly all methods that included protein-evidence (TSB (SR/TRINITY), TSB (SR/OrthoDB)). The BUSCO completeness scores of about 85% post BRAKER are comparable to those from StringTie2.

In the case of *Liriodendron*'s BRAKER and TSEBRA runs, the mono:multi ratios were more variable when compared to the StringTie2 runs, which ranged from 0.34 BR(SR) to 1.04 with BR (SR/RM2+). The annotation rates for each run were around 75%, with BUSCO scores from 83% with TSB (SR/LR/ST2) to 90.8% for BR (SR). *Populus* gene models post-BRAKER without protein had mono:multi scores around 0.24, and with TSEBRA, the ratio ranged from 0.4 to 0.5. Annotation rates also differed between BR(SR) (75%) and TSB (SR/ST2) (88%). *Rosa* had overall consistent scores for BUSCO post BRAKER, ranging around 96%. TSEBRA runs had higher mono:multi ratios (0.75) compared to the BRAKER runs (0.37) (Table S8).

Annotation with long reads-

For BRAKER runs, the predicted gene lengths from the long-reads were comparable to those based on short-reads, except for some runs with long-reads within *Populus*. The average gene length for BRAKER with long-reads for *Populus* ranges from 2.7 Kb to 3.4 Kb, although some transcripts exceed 6 Kb. The longest average gene lengths were observed in *Liriodendron* (9.3 Kb- BR; SR/LR). The inclusion of long-reads (exclusively) did not improve BUSCO completeness for any species, apart from *Arabidopsis*, where the BR (LR) BUSCO

172 completeness was 1% higher than the BR (SR) run. The increase in BUSCO completeness in
173 *Arabidopsis* could be due to the large number of long-reads included (23M across four libraries).
174 However, the quality of genome annotation does not seem correlated with depth of long-read
175 sequencing; for example, *Rosa* had more reads (41M across 6 libraries), and the BR (LR) run
176 had a similar BUSCO score to BR (SR) (96%). It should be noted that the long reads for
177 *Arabidopsis* and *Rosa* were sequenced with ONT. The ONT reads had higher mapping rates,
178 compared to Iso-Seq, to their respective genomes, 97.1% in *Arabidopsis* and 99% in *Rosa*
179 (Table S5). The long-read inputs, regardless of depth or type, lowered the BUSCO
180 completeness (up to 10%) across all ST2 (LR) runs (Table S9). Finally, we note that the
181 combination of short-reads and long-reads (BR; SR/LR and TSB; SR/LR/ST2) is comparable to
182 the BR (SR) reads in terms of BUSCO completeness. Exceptions in this case are *Arabidopsis*
183 and *Rosa*, with marginally higher BUSCO completeness scores. BR (SR/LR) runs produced
184 more genes (in total) for *Arabidopsis*, *Populus*, and *Rosa* in comparison to BR (SR); however, in
185 the TSB (SR/LR/ST2) runs, the total number of genes were lower compared to BR (SR), apart
186 from *Rosa*. The annotation rate in BR (SR/LR) of *Liriodendron* was higher than BR (SR). The
187 combination of proteins with SR and LR (TSB(SR/LR/ST2)) resulted in higher annotation rates
188 across all species. The combination of SR and LR increased mono: multi ratios, therefore,
189 resulted in worse mono: multi ratios across all species.
190

191 ***Refining the genome annotation for Liriodendron-***

192 The BRAKER runs for *Liriodendron* were filtered with gFACs and InterProScan to remove
193 unlikely gene models (Table 4). The number of mono-exonic genes was drastically reduced
194 post-filter with InterProScan. Across all runs, the mono-exonic genes numbered 11K to 25K.
195 After removing mono-exonics without a protein domain annotation from the Pfam database, they
196 decreased from 13K to 5K. The decrease in false positive mono-exonics resulted in an

197 improved mono:multi ratio that range from 0.16 for BR (SR) and BR (SR/RM2+), 0.16 and 0.23
198 for the StringTie2 runs to 0.43 for the TSEBRA runs. The BUSCO scores decreased slightly
199 post-filtering (1-2%). EnTAP annotation rates ranged from 66% to 86%, with the TSB
200 (SR/OrthoDB), which is an improvement from 59% to 68% pre-filtering.

201 In terms of BUSCO completeness and mono:multi ratios, the two best performing runs (BR (SR)
202 and BR (SR/LR)) were further filtered (Table 4). In this step, multi-exonic genes without an
203 EggNOG alignment or a sequence similarity alignment through EnTAP were removed. These
204 filtered models were re-assessed for mono:multi ratio, BUSCO completeness, and EnTAP
205 annotation. The BUSCO completeness remained the same for BR(SR), but not for BR(SR/LR).
206 The EnTAP annotation increased from 66% to 81% in BR (SR), and 67% to 87% in BR (SR/LR).

207 DISCUSSION

208 BRAKER (Hoff et al., 2020) and MAKER (Cantarel et al., 2008) are currently the most popular
209 eukaryotic structural annotation tools, cited 475 and 1,010 times (since 2021, as referenced in
210 Google Scholar). Processes that select from multiple *ab initio* or aligned forms of evidence are
211 gaining popularity as well though they add both time and complexity to the analyses (FINDER
212 cited 22 times, (Banerjee et al., 2021); EVidenceModeler cited 381 times (Haas et al., 2008)).
213 Finally, as high-throughput transcriptomics, in the form of both short and long-read evidence
214 become more accessible, rapid approaches like StringTie2 (cited 451 times (Kovaka et al.,
215 2019)) are occasionally used as the exclusive approach, though more often, used in
216 combination with the options listed above.
217
218 Regardless of the methods selected, recently published benchmarks are challenged to achieve
219 high values for gene sensitivity in larger genomes (Brůna et al., 2021). Within smaller and less
220 complex model systems such as *C. elegans* and *D. melanogaster*, *ab initio* prediction results in

gene sensitivity 49.8% and 59.5%, respectively (Brůna et al., 2021). In well studied complex organisms, such as humans, gene level sensitivity and specificity hovers at 48% and 43%, respectively (Banerjee et al., 2021). While generating benchmarks with model systems (*A. thaliana*, *C. elegans*, and *D. melanogaster*) provides more reliable metrics for comparison, they are infamous for not fully representing the diversity of their respective clades (Chang et al., 2016).

This study focused on four gene prediction workflows: StringTie2, MAKER, BRAKER, and BRAKER/TSEBRA, and examined the process across a variety of evidence inputs. Both model and non-model plant genomes were considered to highlight the challenges and reinforce the need for downstream filtering.

Genome annotation benchmarks for both models and non-models-

Among plant genomes, the total number of genes is relatively conserved and ranges from 20,000 to just over 40,000. As such, total gene number provides an accessible preliminary benchmark. However, the number of genes in the reference annotation fails to assess the overall quality of the annotation. To measure this, we should consider additional metrics. Here, we describe the utility of BUSCO score, mono-exonic: multi-exonic ratio, and sequence similarity assessment.

BUSCO allows us to identify complete, duplicated, fragmented, and missing single-copy orthologs shared by most seed plants (Simão et al. 2015; Seppey, Manni, and Zdobnov 2019). This provides a reliable benchmark in the absence of a high quality reference annotation and poor BUSCO scores are immediately indicative of a larger issue. However, a high BUSCO score is not sufficient to estimate the quality of an annotation (Fig 4B). Six of the 18 BRAKER

246 runs and four of the 17 StringTie2 runs exceeded 95% completeness. However, total gene
247 number, gene length, and structure varied considerably.
248
249 Repeat content, especially in the form of LTRs, and pseudogenes can lead to inflated gene
250 model estimates, especially in the form of mono-exonic genes (Scott et al., 2020; Trouern-Trend
251 et al., 2020). We expect that eukaryotes maintain 20% or less of their gene space as mono-
252 exonics (Jain et al., 2008; Table S3). Although the BUSCO scores were consistent, we note
253 tremendous variation in mono- to multi-gene model ratios post-BRAKER. In practice, having a
254 worse mono: multi ratio is preferable to having a lower BUSCO score, since missing genes,
255 especially those thought to be conserved, cannot be easily rectified, while putative false
256 positives can potentially be filtered through other means.
257
258 Sequence similarity search metrics are more complex to interpret, but when used with high-
259 quality and curated databases that contain full-length proteins (e.g. NCBI RefSeq), can provide
260 a benchmark. Specifically, a reciprocal BLAST search requires that both the query and target in
261 the search retain a minimum level of coverage in the alignment. For new plant genomes, that
262 are in the darkest corners of the tree of life, this might be a less reliable metric. For species that
263 may fare poorly in database comparisons, searches for protein domains can provide some level
264 of confidence and we demonstrate this as a filter to reduce the mono-exonics in *Liriodendron.*
265
266 ***Masking repeats in plant genomes: Repeat masking is important but may not require***
267 ***additional LTR resolution to improve performance-***
268 Plant genomes typically contain many repeats, mostly in the form of transposable elements
269 (TEs), averaging around 46% (Luo et al., 2022). Given the abundance of TEs in genomes, it is
270 important to mask these in advance of gene prediction. Soft masking involves changing
271 nucleotides identified as repeats to lowercase (Yandell & Ence, 2012), signaling downstream

272 programs to ignore these sequences. Of the five genomes included in this study, *Liriodendron*
273 had the largest genome size and repeat content. Running downstream analyses on an
274 unmasked genome of *Liriodendron* resulted in a 4-fold increase in gene predictions (Fig 5A).
275 Many repeats were identified as putative gene models, resulting in a large increase of total
276 number of genes (Fig 5B).
277
278 RepeatModeler2 is a widely used tool for TE discovery (Flynn et al., 2020). The recent release
279 of RepeatModeler2 includes an optional module for more robust LTR structural detection
280 (LTRStruct module) that includes the LTRharvest (Ellinghaus et al., 2008), LTRDetector
281 (Valencia & Girgis, 2019), and LTR_retriever packages (Ou & Jiang, 2018). This is particularly
282 useful in identifying more divergent LTRs in the genome that may exist in fewer copies (Ou &
283 Jiang, 2018; Valencia & Girgis, 2019). Among the default packages included, RepeatScout
284 serves as a fast method to detect young and abundant repeat families in the genome. RECON,
285 on the other hand, is more computationally intensive and is sensitive enough to detect older TE
286 families. The LTRStruct module is run on the unmasked genome to identify LTR families that
287 may be redundant with the families identified by the default package. This creates redundancy
288 that is resolved through clustering with CD-HIT (Flynn et al., 2020).
289
290 In the four species compared, additional repeat masking did not significantly improve gene
291 predictions (Table S9; Fig 4). The mono: multi ratios across species were consistent before and
292 after additional LTR masking (Fig 5A). The BUSCO completeness scores remained relatively
293 the same, with BR (SR/RM2+) being 1% higher than BR (SR) in *Arabidopsis, Funaria* and
294 *Populus.* The marginal improvement observed in these genomes could be related to the
295 structure and type of LTRs, for example, better identification of divergent Ty1-copia elements
296 described in the *Funaria* genome (Kirbis et al., 2022). While we did not include genomes with

297 excessive repeat estimates (>70%), our results indicated that the optional LTRStruct module
298 was not beneficial.
299
300 ***Genome-guided transcriptome assembly for annotation: Transcripts derived directly***
301 ***from alignments are not sufficient to annotate reference genomes-***
302
303 Transcriptome assemblers are designed to work with primarily short RNA-Seq reads to
304 construct full-length transcripts. In the presence of a high-quality reference genome, genome-
305 guided approaches are preferred as the reads are aligned directly to the target genome in
306 advance. Aligned RNA evidence provides resolution on exon boundaries, and aids in the
307 identification of splice variants. *De novo* approaches build graph models directly from the short
308 (or long) reads to generate transcripts. The latter is much more challenging, computationally
309 intensive, and prone to error.
310
311 We compared the accuracy of the annotations produced by StringTie2; *de novo* assembled
312 transcripts with Trinity to annotations produced by BRAKER. The selected packages are top
313 performers when compared in their respective categories of genome-guided and *de novo*
314 transcriptome assembly (Sahraeian et al., 2017; Venturini et al., 2018). As expected, Trinity
315 produced a higher number of transcripts than StringTie2, and BUSCO completeness was
316 consistently lower (Table 3), except for *Liriodendron*. The gene models generated by StringTie2
317 were more numerous than the BRAKER gene models, more than expected for each species. It
318 should be noted, however, that StringTie2 identifies splice variants by generating a splice graph
319 and resolving conflict between multiple potential splice sites (Kovaka et al., 2019), whereas
320 BRAKER trains an internal algorithm GeneMark-ET to find specific genes with complete support
321 among all introns to be further used in training Augustus (Hoff et al., 2019).
322

323 StringTie2 runs resulted in lower BUSCO completeness when compared to BRAKER and/or
324 TSEBRA runs (Fig 4A; Table S11). This outcome is supported by the lack of *ab initio* prediction
325 with genome-guided approaches. Inflated mono-exonic predictions (and lower BUSCO scores)
326 were also observed in the StringTie2 genome annotation of the water strider (*Microvelia*
327 *longipes*) (Toubiana et al., 2021). In our study, *Rosa* ST2 (SR/LR) run was closest with a
328 BUSCO score of 97.2%, BR (SR/LR) of 96.9%, and TSB (SR/ST2) of 98% (Table S9, S10).
329
330 ***Including proteins: Genome annotations are improved for models,and closely related***
331 ***species, when full-length proteins sourced from OrthoDB are used in combination with***
332 ***read data.***
333
334 The performance of StringTie2 and Trinity-derived protein evidence was assessed on the
335 predicted gene models using BRAKER and TSEBRA. In this context, the genome-guided or *de*
336 *novo* assembled transcripts were translated into proteins and provided as evidence to train the
337 *ab initio* component of the pipelines. Adding protein evidence to genome annotation can target
338 protein-coding genes leading to more accurate predictions than RNA-Seq evidence alone
339 (Bruna, 2022). This study specifically focused on using protein evidence derived in some
340 fashion from the transcriptomic inputs but also evaluated the recommendation to include clade
341 specific OrthoDB protein inputs to the BRAKER/TSEBRA approach.
342
343 The TSEBRA runs of the model species, *Arabidopsis* and *Populus* were compared to the
344 reference annotations. These runs were the best for the model species in terms of sensitivity
345 and specificity as compared to the MAKER, StringTie2, Trinity and BRAKER runs (Fig 3B). The
346 model genomes also had very similar BUSCO completeness scores but had different mono:
347 multi ratios with the addition of protein evidence. As expected, the model genomes benefitted
348 the most from the inclusion of the external OrthoDB proteins in terms of annotation rate and

349 BUSCO score (both the *Arabidopsis* and *Populus* reference proteins are contained within this
350 resource). However, mono: multi ratio challenges remained consistent across the TSEBRA runs
351 with varying inputs.
352
353 In the case of non-model plant genomes, TSEBRA contributed to higher mono:multi ratios, and
354 this was very evident in *Liriodendron*. However, the BUSCO scores of the non-protein runs were
355 lower across all runs of *Liriodendron*, compared to the protein runs. The *Rosa* TSB
356 (SR/OrthoDB) reported the highest BUSCO score across all runs; given the phylogenetic
357 placement of *Rosa* in comparison to *Arabidopsis*, we believe that the addition of OrthoDB
358 greatly influenced this result. On the other hand, the annotation rate of TSB (SR/OrthoDB)
359 remained like the other runs within *Rosa*. The higher quality of the *Rosa* genome assembly,
360 compared to the other two non-models, could also influence the utility of the protein evidence.
361 However, mono: multi ratios remained to be high, and annotation rates were like those of the
362 runs without OrthoDB proteins.
363
364 TSEBRA runs with proteins sourced from genome-guided predictions perform similarly, but had
365 lower BUSCO, higher mono: multi ratios, and total gene number when compared to the short-
366 read only runs (Fig 4A, Table S9). Among TSEBRA runs, Trinity fares better only for
367 *Liriodendron*, which could indicate that genome-guided proteins are not a suitable choice for a
368 more repetitive genome. This is consistent in independent assessments between *de novo*
369 transcriptome assemblers and genome-guided assemblers with complex genomes with
370 fragmented genome assemblies (for example in *Ae. albopictus* (Huang et al., 2016). TSEBRA
371 with full length proteins sourced from OrthoDB had lower BUSCO scores when compared to
372 BR(SR) for the non-model species, *Liriodendron* and *Funaria.* In both cases, TSB
373 (SR/OrthoDB).
374

375 The total number of genes predicted by TSEBRA and BRAKER runs remained largely the same
376 across all species (Table S9). However, the number of mono-exonic genes increased, whereas
377 the multi-exonic genes decreased across all TSEBRA runs in comparison to the BRAKER runs
378 without proteins across all species. The average gene lengths also decreased, while average
379 lengths of mono-exonics remained the same, and the lengths of multi-exonics are higher.
380
381 Initial examination of the EnTAP reciprocal BLAST assessment revealed high annotation rates
382 for the non-model species when protein evidence was included, particularly the multi-exonics
383 (whereas the mono-exonic percentage remained the same) (Table S9). However, this increase
384 in multi-exonic annotation proved to be an artifact since the total number of multi-exonic models
385 was reduced. Direct comparison of the predictions revealed that 40%-52% of the multi-exonics
386 were split into mono-exonic predictions when comparing the BR(SR) to the TSB (SR/ST2) and
387 TSB (SR/OrthoDB) gene models predicted using *Liriodendron* (Table S14).
388
389 ***Long-read transcriptomes: Long-reads can be paired with short-reads to improve the***
390 ***quality of the resulting models-***
391 Long-reads generated from platforms such as Oxford Nanopore or PacBio have the potential to
392 resolve splice variants and assemble transcripts more accurately than traditional Illumina RNA-
393 Seq (Amarasinghe et al., 2020). While long-reads can independently generate transcriptomes, it
394 is recommended to have a combination of short and long reads to achieve greater depth,
395 improved error profiles, and gain more evidence for splice site resolution (Amarasinghe et al.,
396 2020; Gonzalez-Ibeas et al., 2016; Watson & Warr, 2019).
397
398 In this study, we utilized both ONT and Iso-Seq long-reads. In the latter, we relied on raw reads
399 (not the error-corrected CCS reads) in our comparisons for genome annotations using long-
400 reads. In all cases, long-reads (alone) did not outperform short-reads for the BRAKER runs.

401 However, in some cases, the combination of short-read and long-read inputs was beneficial.
402 The higher error rate Iso-Seq reads from *Populus* and *Liriodendron* produced comparable, but
403 lower, BUSCO scores compared to the BR (SR) runs. In contrast, the ONT long-reads used for
404 *Arabidopsis* and *Rosa* in the combined runs (BR (SR/LR)) had slightly better BUSCO
405 completeness as compared to the BR (SR) runs, and similar mono:multi ratios. Overall, the
406 lower error profile of using ONT reads, supplemented with short-read data, as well as using high
407 quality reference genomes, support the higher BUSCO completeness scores.
408
409 ***Best Practices for Plant Genome Annotation-***
410 Given existing tools, we recommend that investigators utilize RepeatModeler2 to mask their
411 genome of interest with the default settings (Flynn et al. 2020). Following soft-masking, RNA-
412 Seq short reads (between 4-10 libraries, paired-end, minimum 15M reads per library) are
413 generally sufficient for annotation. While we did not comprehensively investigate the impact of
414 tissue type, it is recommended to sample from multiple tissues when possible (Kress et al.,
415 2022). In our study, we did not observe a difference in the annotation completeness among
416 species with a higher number of short-read libraries, although we did not comprehensively
417 evaluate the difference of using fewer libraries within a single species.
418
419 Sequencing of long-read libraries remains more expensive than generating deep Illumina short-
420 read RNA-Seq. In most cases, the short reads were sufficient as input. The notable exceptions
421 include the BR (SR/LR), as they were comparable, and in some cases, slightly improved, over
422 BR (SR) across all species. The lower error-profile Nanopore reads were more beneficial when
423 combined with short-reads. However, current long-read technologies available from both
424 platforms may provide different results.
425

426 BRAKER and TSEBRA outperformed runs of MAKER, StringTie2 and Trinity with default
427 settings. It should be noted that the authors did not comprehensively benchmark MAKER with
428 more than two training runs of AUGUSTUS as recommended, which could have further
429 improved results. However, previous benchmarking studies also support lower performance of
430 MAKER (Banerjee et al., 2021; Hoff et al., 2020). Among the BRAKER runs executed in the
431 model plants, *Arabidopsis* and *Populus*, the TSEBRA runs were the best runs. TSEBRA,
432 especially TSB (SR/OrthoDB), performs the best for *Rosa* but would require substantial filtering
433 to remove false positives. Among the less contiguous and more evolutionary distant species
434 (*Funaria* and *Liriodendron*), BR (SR) runs performed the best in terms of BUSCO completeness
435 and mono:multi ratios. Overall EnTAP annotation rates were greatly improved in runs where
436 OrthoDB proteins were included as evidence. However, when considering BUSCO as well as
437 mono: multi ratio, especially for non-model species, the BR(SR) runs performed the best. For
438 more divergent species (as defined by current public databases), BRAKER runs with short-
439 reads, or short-reads and long-reads, are advised.
440
441 Regardless of approach, existing pipelines do not provide appropriate summary statistics or
442 robust methods for filtering unlikely gene models. All methods produce more putative false
443 positives than desired. We recommend utilizing reciprocal BLAST searches with well curated
444 databases containing targets with full-length proteins (such as NCBI's RefSeq) to identify
445 fragmented models. We also recommend filtering and removing mono-exonics that do not have
446 a protein domain. Finally, we recommend structural filters to remove unlikely gene structures
447 (splice sites, start sites, incompletes, etc).
448
449 In this study, we demonstrated the impact of post-filtering on the most complex genome
450 assessed in this study, *Liriodendron*. We improved the published annotation across all
451 benchmarks evaluated in this study following a new BR(SR) run (Table 4; (Chen et al., 2019)).

The filters reduced the overall number of putative false positives and increased the overall rate of annotation, with minimal reduction to BUSCO completeness.

Data Availability: All scripts and data used is available through https://www.protocols.io/blind/3A33C8E3B76511EC84CA0A58A9FEAC02 . The public data (NCBI SRA and genome assembly accessions) for the reference genomes, short-reads, and long-reads are listed in Table S2.

Acknowledgements

The authors would like to thank the Institute for Systems Genomics (ISG) and Computational Biology Core at UConn for HPC services. Thanks go to Dr. Bernard Goffinet for his careful review of this manuscript. This work was supported by the National Science Foundation awards, DEB-1753811 and DBI-1943371.

1 LITERATURE CITED

2 Amarasinghe, S. L., Su, S., Dong, X., Zappia, L., Ritchie, M. E., & Gouil, Q. (2020).
3 Opportunities and challenges in long-read sequencing data analysis. *Genome Biology*,
4 *21*(1), 30.
5 Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available
6 online. *Retrieved May*, *17*, 2018.
7 Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant
8 Arabidopsis thaliana. *Nature*, *408*(6814), 796–815.
9 Armisén, D., Lecharny, A., & Aubourg, S. (2008). Unique genes in plants: specificities and
10 conserved features throughout evolution. *BMC Evolutionary Biology*, *8*, 280.
11 Banerjee, S., Bhandary, P., Woodhouse, M., Sen, T. Z., Wise, R. P., & Andorf, C. M. (2021).
12 FINDER: an automated software package to annotate eukaryotic genes from RNA-Seq
13 data and associated protein sequences. *BMC Bioinformatics*, *22*(1), 205.
14 Bolger, M. E., Arsova, B., & Usadel, B. (2018). Plant genome and transcriptome annotations:
15 from misconceptions to simple solutions. *Briefings in Bioinformatics*, *19*(3), 437–449.
16 Bruna, T. (2022). *Unsupervised algorithms for automated gene prediction in novel eukaryotic*
17 *genomes*. https://smartech.gatech.edu/handle/1853/67297
18 Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M., & Borodovsky, M. (2021). BRAKER2:
19 automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported
20 by a protein database. *NAR Genomics and Bioinformatics*, *3*(1), lqaa108.
21 Caballero, M., & Wegrzyn, J. (2019). gFACs: Gene Filtering, Analysis, and Conversion to Unify
22 Genome Annotations Across Alignment and Gene Prediction Frameworks. *Genomics,*
23 *Proteomics & Bioinformatics*, *17*(3), 305–310.
24 Campbell, M. S., Holt, C., Moore, B., & Yandell, M. (2014). Genome Annotation and Curation

25 Using MAKER and MAKER-P. *Current Protocols in Bioinformatics / Editoral Board,*
26 *Andreas D. Baxevanis ... [et Al.]*, *48*, 4.11.1–39.
27 Campbell, M. S., Law, M., Holt, C., Stein, J. C., Moghe, G. D., Hufnagel, D. E., Lei, J.,
28 Achawanantakun, R., Jiao, D., Lawrence, C. J., Ware, D., Shiu, S.-H., Childs, K. L., Sun,
29 Y., Jiang, N., & Yandell, M. (2014). MAKER-P: A Tool Kit for the Rapid Creation,
30 Management, and Quality Control of Plant Genome Annotations. *Plant Physiology*, *164*(2),
31 513–524.
32 Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez
33 Alvarado, A., & Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed
34 for emerging model organism genomes. *Genome Research*, *18*(1), 188–196.
35 Chang, C., Bowman, J. L., & Meyerowitz, E. M. (2016). Field Guide to Plant Model Systems.
36 *Cell*, *167*(2), 325–339.
37 Cheng, S., Melkonian, M., Smith, S. A., Brockington, S., Archibald, J. M., Delaux, P.-M., Li, F.-
38 W., Melkonian, B., Mavrodiev, E. V., Sun, W., Fu, Y., Yang, H., Soltis, D. E., Graham, S.
39 W., Soltis, P. S., Liu, X., Xu, X., & Wong, G. K.-S. (2018). 10KP: A phylodiverse genome
40 sequencing plan. *GigaScience*, *7*(3), 1–9.
41 Chen, J., Hao, Z., Guang, X., Zhao, C., Wang, P., Xue, L., Zhu, Q., Yang, L., Sheng, Y., Zhou,
42 Y., Xu, H., Xie, H., Long, X., Zhang, J., Wang, Z., Shi, M., Lu, Y., Liu, S., Guan, L., … Shi,
43 J. (2019). Liriodendron genome sheds light on angiosperm phylogeny and species-pair
44 differentiation. *Nature Plants*, *5*(1), 18–25.
45 Corchete, L. A., Rojas, E. A., Alonso-López, D., De Las Rivas, J., Gutiérrez, N. C., & Burguillo,
46 F. J. (2020). Systematic comparison and assessment of RNA-seq procedures for gene
47 expression quantitative analysis. *Scientific Reports*, *10*(1), 19737.
48 Deutekom, E. S., Vosseberg, J., van Dam, T. J. P., & Snel, B. (2019). Measuring the impact of
49 gene prediction on gene loss estimates in Eukaryotes by quantifying falsely inferred
50 absences. *PLoS Computational Biology*, *15*(8), e1007301.

51 Edgar, R. (2010). *Usearch*. https://www.osti.gov/biblio/1137186

52 Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software
53 for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, *9*, 18.

54 Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F.
55 (2020). RepeatModeler2 for automated genomic discovery of transposable element
56 families. *Proceedings of the National Academy of Sciences of the United States of America*,
57 *117*(17), 9451–9457.

58 Gabriel, L., Hoff, K. J., Brůna, T., Borodovsky, M., & Stanke, M. (2021). TSEBRA: transcript
59 selector for BRAKER. *BMC Bioinformatics*, *22*(1), 566.

60 Gonzalez-Ibeas, D., Martinez-Garcia, P. J., Famula, R. A., Delfino-Mix, A., Stevens, K. A.,
61 Loopstra, C. A., Langley, C. H., Neale, D. B., & Wegrzyn, J. L. (2016). Assessing the Gene
62 Content of the Megagenome: Sugar Pine (Pinus lambertiana). *G3* , *6*(12), 3787–3802.

63 Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X.,
64 Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A.,
65 Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011).
66 Full-length transcriptome assembly from RNA-Seq data without a reference genome.
67 *Nature Biotechnology*, *29*(7), 644–652.

68 Gremme, G. (2014). *GenomeThreader Gene Prediction Software*.
69 https://genomethreader.org/doc/gthmanual.pdf

70 Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for
71 genome assemblies. *Bioinformatics* , *29*(8), 1072–1075.

72 Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., &
73 Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using
74 EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*,
75 *9*(1), R7.

76 Hart, A. J., Ginzburg, S., Xu, M. S., Fisher, C. R., Rahmatpour, N., Mitton, J. B., Paul, R., &

77 Wegrzyn, J. L. (2020). EnTAP: Bringing faster and smarter functional annotation to non-
78 model eukaryotic transcriptomes. *Molecular Ecology Resources*, *20*(2), 591–604.
79 Hoff, K. J., Brŭna, T., Lomsadze, A., & Stanke, M. (2020). Fully Automated and Accurate
80 Annotation of Eukaryotic Genomes with BRAKER2. *Poster Presented at*.
81 https://www.researchgate.net/profile/Katharina-Hoff-
82 2/publication/338831355_Fully_Automated_and_Accurate_Annotation_of_Eukaryotic_Gen
83 omes_with_BRAKER2/links/5e2d9102299bf152167f6424/Fully-Automated-and-Accurate-
84 Annotation-of-Eukaryotic-Genomes-with-BRAKER2.pdf
85 Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., & Stanke, M. (2016). BRAKER1:
86 Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS.
87 *Bioinformatics* , *32*(5), 767–769.
88 Hoff, K. J., Lomsadze, A., Borodovsky, M., & Stanke, M. (2019). Whole-Genome Annotation
89 with BRAKER. *Methods in Molecular Biology* , *1962*, 65–95.
90 Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database
91 management tool for second-generation genome projects. *BMC Bioinformatics*, *12*, 491.
92 Huang, X., Chen, X.-G., & Armbruster, P. A. (2016). Comparative performance of transcriptome
93 assembly methods for non-model organisms. *BMC Genomics*, *17*, 523.
94 Jain, M., Khurana, P., Tyagi, A. K., & Khurana, J. P. (2008). Genome-wide analysis of intronless
95 genes in rice and Arabidopsis. *Functional & Integrative Genomics*, *8*(1), 69–78.
96 Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J.,
97 Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M.,
98 Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function
99 classification. *Bioinformatics* , *30*(9), 1236–1240.
100 Joshi NA, F. J. N. (2011). *Sickle: A sliding-window, adaptive, quality-based trimming tool for*
101 *FastQ files* (Version Version 1.33) [Computer software]. https://github.com/najoshi/sickle
102 Jung, H., Ventura, T., Chung, J. S., Kim, W.-J., Nam, B.-H., Kong, H. J., Kim, Y.-O., Jeon, M.-

S., & Eyun, S.-I. (2020). Twelve quick steps for genome assembly and annotation in the classroom. *PLoS Computational Biology*, *16*(11), e1008325.

Kersey, P. J. (2019). Plant genome sequences: past, present, future. *Current Opinion in Plant Biology*, *48*, 1–8.

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

Kirbis, A., Rahmatpour, N., Dong, S., Yu, J., van Gessel, N., Waller, M., Reski, R., Lang, D., Rensing, S. A., Temsch, E. M., Wegrzyn, J. L., Goffinet, B., Liu, Y., & Szövényi, P. (2022). Genome dynamics in mosses: Extensive synteny coexists with a highly dynamic gene space. In *bioRxiv* (p. 2022.05.17.492078). https://doi.org/10.1101/2022.05.17.492078

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*(1), 59.

Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg, S. L., & Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, *20*(1), 278.

Kress, W. J., Soltis, D. E., Kersey, P. J., Wegrzyn, J. L., Leebens-Mack, J. H., Gostel, M. R., Liu, X., & Soltis, P. S. (2022). Green plant genomes: What we know in an era of rapidly expanding opportunities. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(4). https://doi.org/10.1073/pnas.2115640118

Lewin, H. A., Richards, S., Lieberman Aiden, E., Allende, M. L., Archibald, J. M., Bálint, M., Barker, K. B., Baumgartner, B., Belov, K., Bertorelle, G., Blaxter, M. L., Cai, J., Caperello, N. D., Carlson, K., Castilla-Rubio, J. C., Chaw, S.-M., Chen, L., Childers, A. K., Coddington, J. A., … Zhang, G. (2022). The Earth BioGenome Project 2020: Starting the clock. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(4). https://doi.org/10.1073/pnas.2115635118

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* , *34*(18),

3094–3100.

Li, H. (2021). New strategies to improve minimap2 alignment accuracy. *Bioinformatics* . https://doi.org/10.1093/bioinformatics/btab705

Luo, X., Chen, S., & Zhang, Y. (2022). PlantRep: a database of plant repetitive elements. *Plant Cell Reports*, *41*(4), 1163–1166.

Manni, M., Berkeley, M. R., Seppey, M., Simão, F. A., & Zdobnov, E. M. (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Molecular Biology and Evolution*, *38*(10), 4647–4654.

Marks, R. A., Hotaling, S., Frandsen, P. B., & VanBuren, R. (2021). Representation and participation across 20 years of plant genome sequencing. *Nature Plants*, *7*(12), 1571–1578.

Meyer, C., Scalzitti, N., Jeannin-Girardon, A., Collet, P., Poch, O., & Thompson, J. D. (2020). Understanding the causes of errors in eukaryotic protein-coding gene prediction: a case study of primate proteomes. *BMC Bioinformatics*, *21*(1), 513.

Mudge, J. M., & Harrow, J. (2016). The state of play in higher eukaryote gene annotation. *Nature Reviews. Genetics*, *17*(12), 758–772.

Musich, R., Cadle-Davidson, L., & Osier, M. V. (2021). Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider. *Frontiers in Plant Science*, *12*, 657240.

Ou, S., & Jiang, N. (2018). LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiology*, *176*(2), 1410–1422.

Pertea, G., & Pertea, M. (2020). GFF Utilities: GffRead and GffCompare. *F1000Research*, *9*. https://doi.org/10.12688/f1000research.23297.2

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L.

(2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295.

Pucker, B., Irisarri, I., de Vries, J., & Xu, B. (2022). Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. *Quantitative Plant Biology*, *3*, e5.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., & Lopez, R. (2005). InterProScan: protein domains identifier. *Nucleic Acids Research*, *33*(Web Server issue), W116–W120.

Sahraeian, S. M. E., Mohiyuddin, M., Sebra, R., Tilgner, H., Afshar, P. T., Au, K. F., Bani Asadi, N., Gerstein, M. B., Wong, W. H., Snyder, M. P., Schadt, E., & Lam, H. Y. K. (2017). Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. *Nature Communications*, *8*(1), 59.

Salzberg, S. L. (2019). Next-generation genome annotation: we still struggle to get it right. *Genome Biology*, *20*(1), 92.

Scott, A. D., Zimin, A. V., Puiu, D., Workman, R., Britton, M., Zaman, S., Caballero, M., Read, A. C., Bogdanove, A. J., Burns, E., Wegrzyn, J., Timp, W., Salzberg, S. L., & Neale, D. B. (2020). A Reference Genome Sequence for Giant Sequoia. *G3* , *10*(11), 3907–3919.

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.

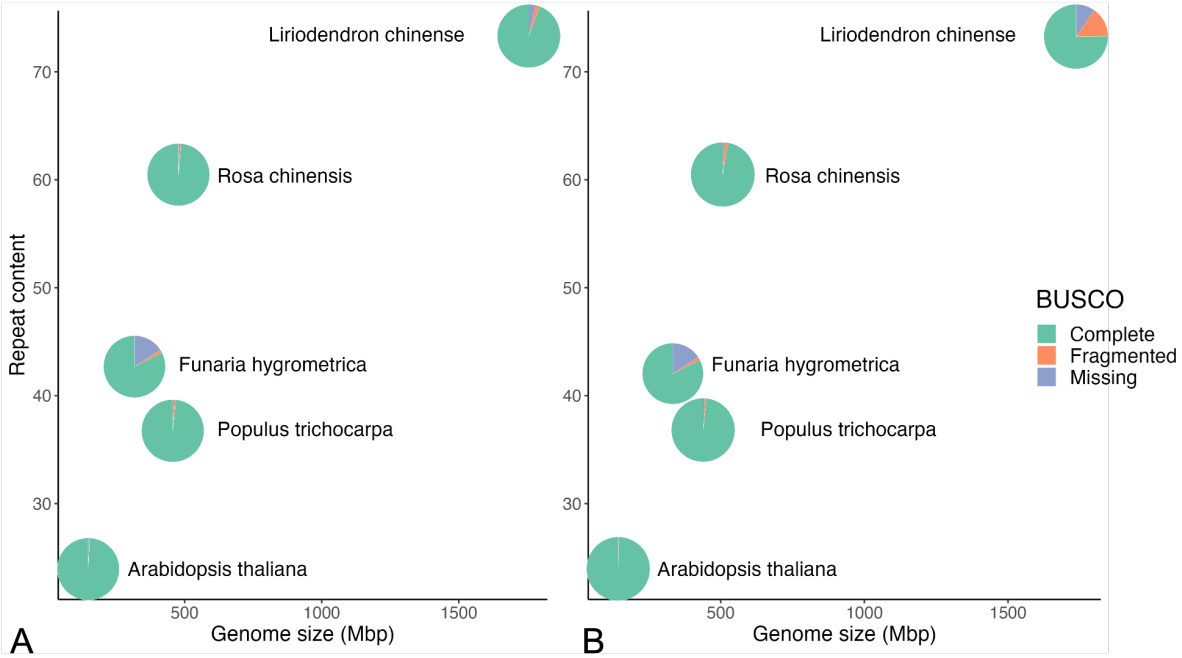
Smit, AFA, Hubley, R & Green, P. (2013-2015). *RepeatMasker Open-4.0*. RepearMasker. http://www.repeatmasker.org

Stanke, M., & Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* , *19 Suppl 2*, ii215–ii225.

Sun, Y., Shang, L., Zhu, Q.-H., Fan, L., & Guo, L. (2022). Twenty years of plant genome sequencing: achievements and challenges. In *Trends in Plant Science* (Vol. 27, Issue 4, pp. 391–401). https://doi.org/10.1016/j.tplants.2021.10.006

181 Toubiana, W., Armisén, D., Dechaud, C., Arbore, R., & Khila, A. (2021). Impact of male trait
182 exaggeration on sex-biased gene expression and genome architecture in a water strider.
183 *BMC Biology*, *19*(1), 89.
184 Trouern-Trend, A. J., Falk, T., Zaman, S., Caballero, M., Neale, D. B., Langley, C. H.,
185 Dandekar, A. M., Stevens, K. A., & Wegrzyn, J. L. (2020). Comparative genomics of six
186 Juglans species reveals disease-associated gene family contractions. *The Plant Journal:*
187 *For Cell and Molecular Biology*, *102*(2), 410–423.
188 Valencia, J. D., & Girgis, H. Z. (2019). LtrDetector: A tool-suite for detecting long terminal repeat
189 retrotransposons de-novo. *BMC Genomics*, *20*(1), 450.
190 Venturini, L., Caim, S., Kaithakottil, G. G., Mapleson, D. L., & Swarbreck, D. (2018). Leveraging
191 multiple transcriptome assembly methods for improved gene structure annotation.
192 *GigaScience*, *7*(8). https://doi.org/10.1093/gigascience/giy093
193 Watson, M., & Warr, A. (2019). Errors in long-read assemblies can critically affect protein
194 prediction [Review of *Errors in long-read assemblies can critically affect protein prediction*].
195 *Nature Biotechnology*, *37*(2), 124–126.
196 Wu, T. D., & Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for
197 mRNA and EST sequences. *Bioinformatics* , *21*(9), 1859–1875.
198 Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature*
199 *Reviews. Genetics*, *13*(5), 329–342.

200 FIGURES



201

202 *Figure 1: Genome size, repeat content, and BUSCO completeness for the five plant genomes:*
203 *Arabidopsis, Populus, Funaria, Rosa, and Liriodendron. Each pie represents the BUSCO*
204 *completeness. Green denotes the completeness score, orange indicates the fragmented score,*
205 *and blue indicates the missing score from BUSCO.* (A) *BUSCO scores estimated from the*
206 ***published assemblies.*** *(B) BUSCO scores estimated from protein-coding gene predictions from*
207 *the* ***published annotations****.*

208

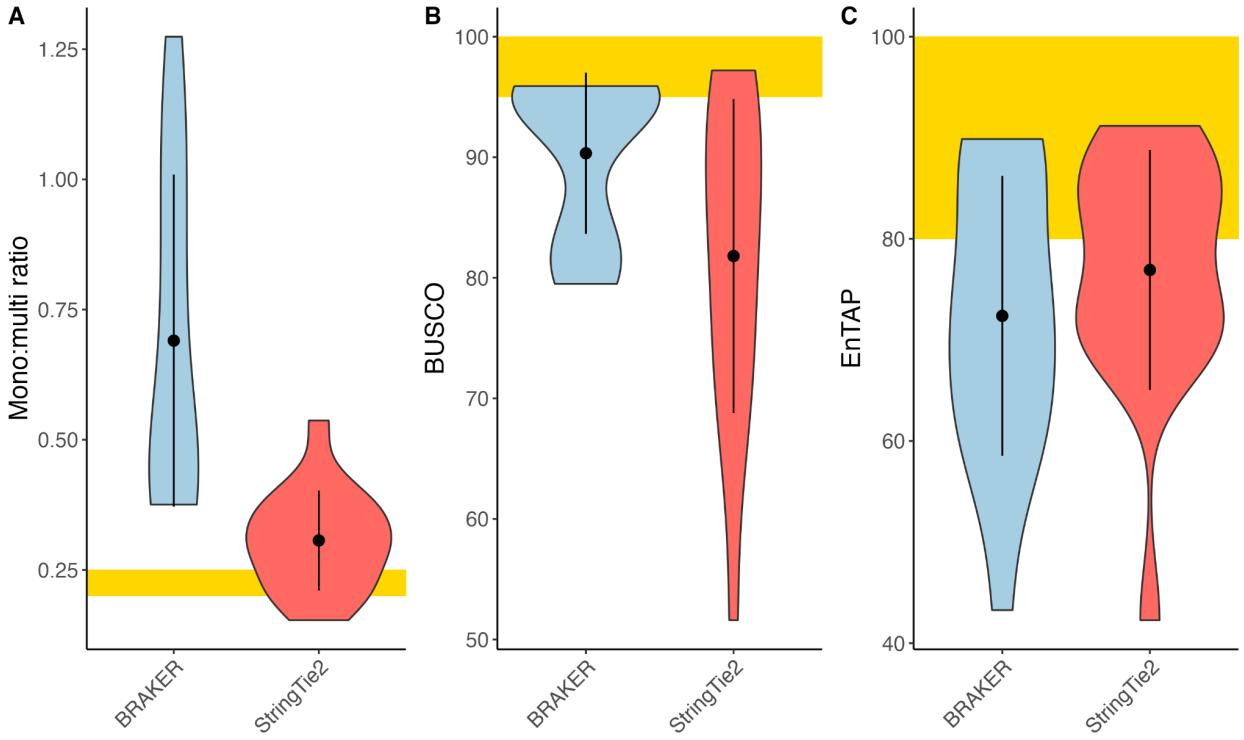
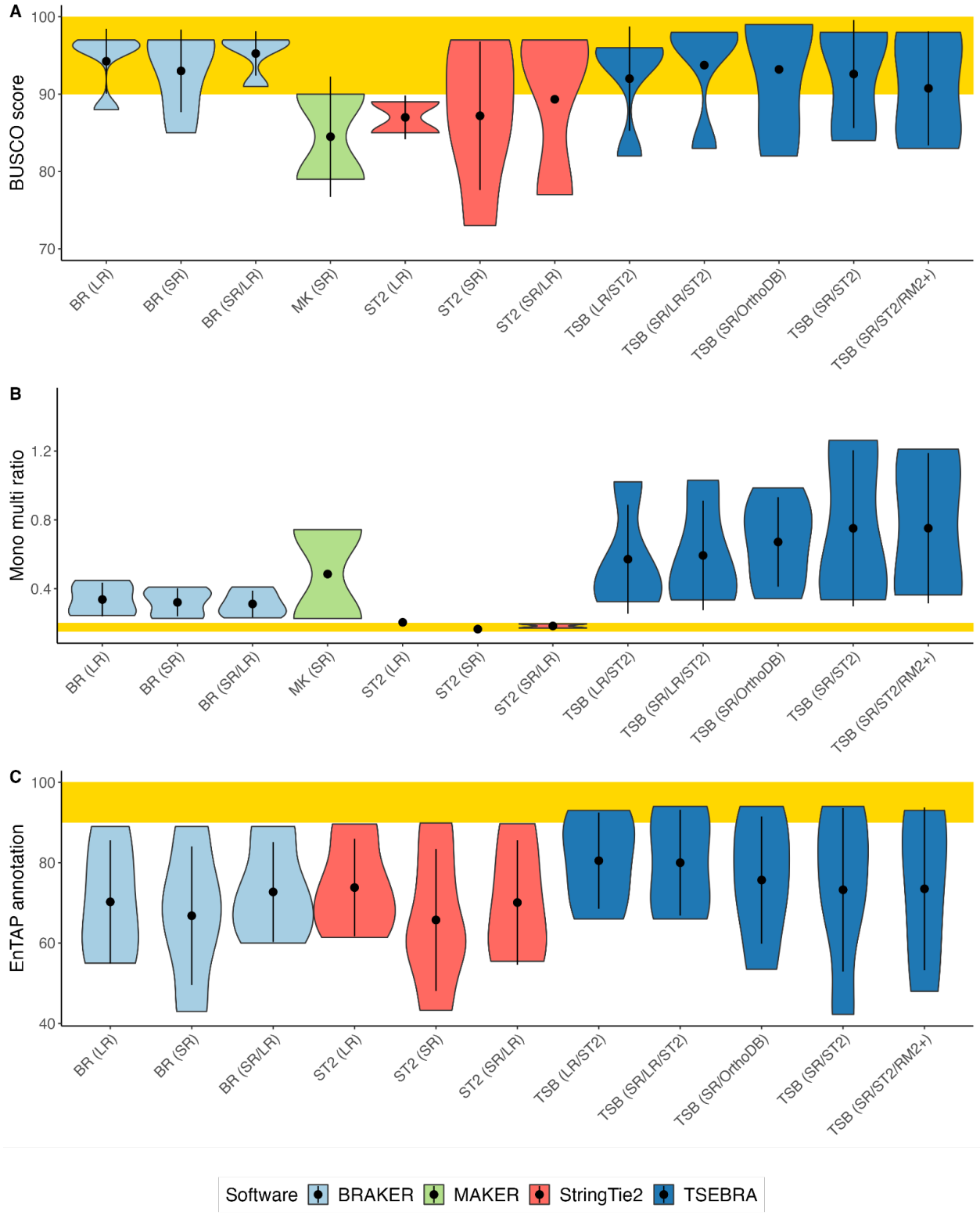


Figure 2: Comparing metrics between BRAKER (blue) and StringTie2 (red) predictions. (A) mono:multi ratios, (B) BUSCO comparisons, and (C) EnTAP annotation rates of the gene models. The yellow region indicates the ideal value for each of the metrics.

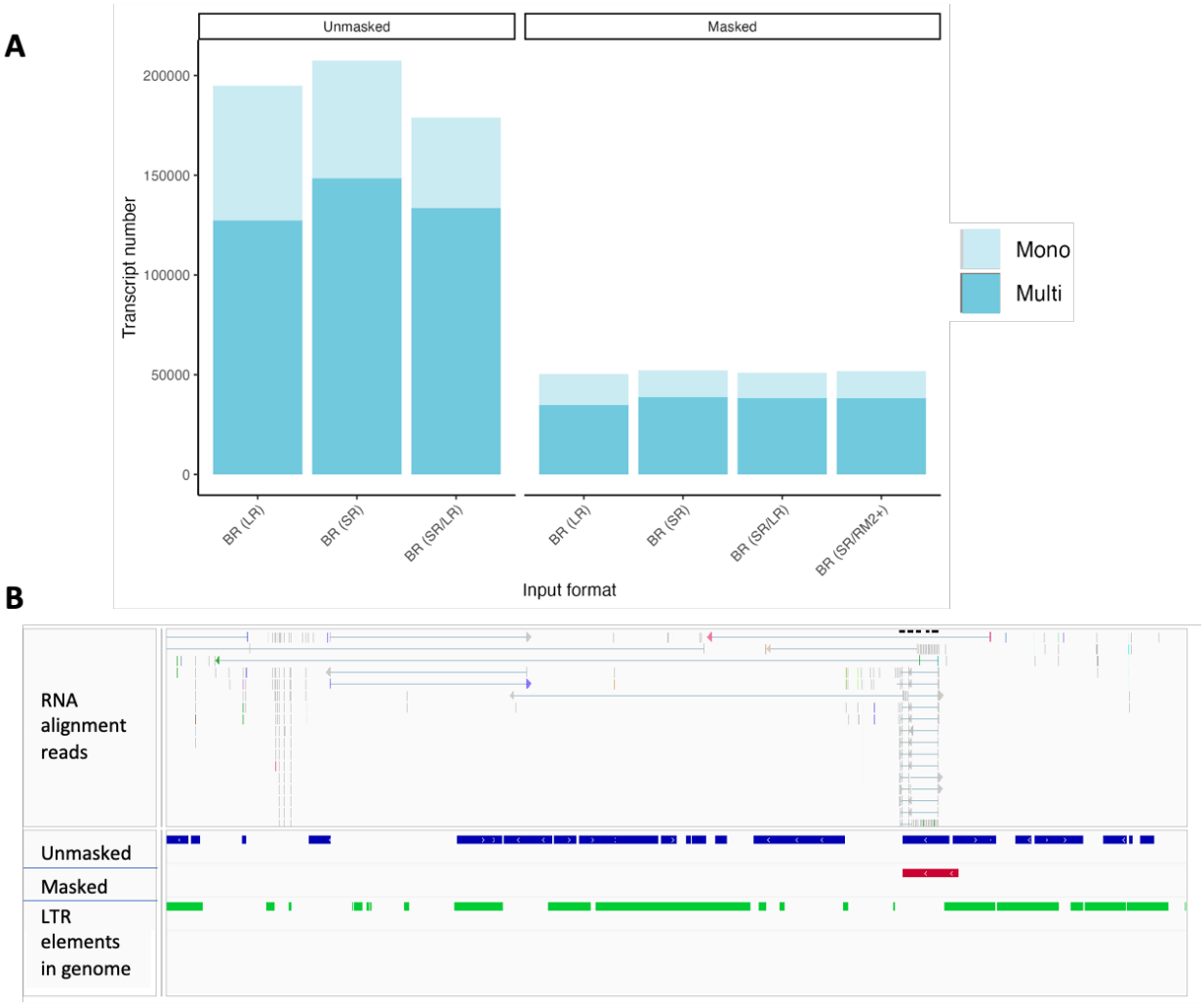
Figure 3: Comparison of BUSCO, sensitivity, and false positive rates between the Arabidopsis and Populus annotations (Table S10). (A) BUSCO completeness scores for the MK(SR/RM2+) and BR(SR/RM2+) runs of Arabidopsis and Populus, green denotes the completeness score, orange indicates the fragmented score, and blue indicates the missing score from BUSCO (B) False positive rates and sensitivity scores from Mikado against published annotations for Arabidopsis (red color) and Populus (gold color) for the MAKER, BRAKER, TSEBRA, Trinity, and StringTie2 runs. The scores were assessed using MIKADO. Multiple points per run reflect differences in input read type and repeat-masking.


A
BUSCO score
100
90
80
70
BR (LR)
BR (SR)
BR (SR/LR)
MK (SR)
ST2 (LR)
ST2 (SR)
ST2 (SR/LR)
TSB (LR/ST2)
TSB (SR/LR/ST2)
TSB (SR/OrthoDB)
TSB (SR/ST2)
TSB (SR/ST2/RM2+)
B
Mono multi ratio
1.2
0.8
0.4
C
EnTAP annotation
100
80
60
40
Software
BRAKER
MAKER
StringTie2
TSEBRA

226

227

228 *Figure 4: Comparison of BUSCO completeness scores (A), mono:multi ratios (B), and EnTAP*
229 *annotation rates (C) across all species between the runs of different input types and software,*
230 *i.e., MAKER (MK-green) BRAKER (BK-light blue),TSEBRA (TSB-dark blue) and StringTie2 (ST2-*
231 *red). The yellow rectangle represents the target scores for each benchmark. RM2+-*
232 *RepeatModeler2 with LTRStruct.*



233

234

235 *Figure 5: (A) The effect of soft-masking on gene prediction in Liriodendron (Table S13).*
236 *Performing structural annotation on the unmasked Liriodendron genome results in inflation in the*
237 *mono and multi- exonic genes. Blue denotes the BRAKER (BR) runs for both genomes, SR*
238 *denotes short-reads, and LR denotes long reads. The lighter shade represents mono-exonics,*

239 *and the darker shade represents the multi-exonics. (B) More genes predicted using the unmasked*
240 *genome (blue), as compared to only one gene predicted in this region with the masked genome*
241 *(red). The green track shows the LTR elements in the genome as identified by RepeatModeler2.*
242 *The RNA alignment reads show a read pile-up at the predicted gene (masked track).*

243 TABLES

244 Table 1: Notations for the different runs performed for benchmarking.

Run		*Arabidopsis*	*Funaria*	*Populus*	*Liriodendron*	*Rosa*
StringTie2						
ST2 (SR)	Short-reads	X	X	X	X	X
ST2 (LR)	Long-reads	X		X	X	X
ST2(SR/LR)	Short and long-reads	X		X	X	X
BRAKER						
BR (SR)	Short-reads	X	X	X	X	X
BR (LR)	Long-reads	X		X	X	X
BR (SR/LR)	Short and long-reads	X		X	X	X
BR (SR/RM2+)	Short-reads with additional masking for LTRs	X	X	X	X	
TSEBRA						
TSB (SR/TRINITY)	Short-reads and *de novo* proteins	X	X	X	X	X
TSB (SR/ST2)	Short-reads and genome-guided proteins	X	X	X	X	X
TSB (LR/ST2)	Long-reads and genome-guided proteins	X		X	X	X
TSB (SR/LR/ST2)	Short and long-reads and genome-guided proteins	X		X	X	X
TSB (SR/ST2/RM2+)	Short-reads and genome-guided proteins with additional masking for LTRs	X	X	X	X	
TSB (SR/OrthoDB)	Short-reads and OrthoDB proteins	X	X	X	X	X
MAKER						
MK (RM2+)	Short-reads with additional masking for LTRs	X	X	X		

245 SR- Short reads, LR- Long reads, and RM2+- RepeatModeler2 with the additional repeat masking.

246

247 Table 2: Genome assembly and annotation statistics for the five published plant genomes

Species	Genome size	Total scaffolds (chromosomes)	N50	Repeat content	BUSCO Completeness	
					Genome	Annotation
Arabidopsis thaliana	119 Mb	7 (5)	23.46 Mb	23.6%	99.30%	99.60%
Funaria hygrometrica	327 Mb	687 (26)	1.48 Mb	42.35%	85.60%	86.60%
Liriodendron chinense	1,742 Mb	3,711 (21)	3.53 Mb	73.18%	98.60%	75.10%
Populus trichocarpa	434 Mb	1,446 (19)	19.47 Mb	35.90%	98.80%	98.30%
Rosa chinensis	515 Mb	55 (7)	69.64 Mb	60.53%	98.80%	97.30%

Table 3: Comparison between genome-guided (StringTie2- ST2) and de novo (Trinity) genome annotations.

Species	RM %	RM2+ %	Total short-reads (total libraries)	Total long-reads (total libraries)	Total Trinity transcripts (N50)	Total ST2 transcripts (SR) (N50)	Total ST2 transcripts (LR) (N50)	Total ST2 transcripts (SR/LR) (N50)	BUSCO transcript alignments (%)		
									Trinity (SR)	ST2 (SR)	ST2 (SR/LR)
Arabidopsis	15.2	16.5	511,277,126 (9)	23,134,068 (4)	319,434 (2726)	37,747 (2538)	36,241 (1599)	42,265 (1363)	82.7	95.5	93.6
Funaria	42.3	43.1	549,205,030 (9)		151,265 (1198)	59,741 (369)			72.9	84.5	
Liriodendron	73.2	72.7	1,408,831,670 (20)	10,437,029 (1)	2,839,867 (3055)	62,341 (1041)	33,895 (1815)	45,785 (2361)	92.5	87.1	77.3
Populus	35.9	45.1	267,403,772 (5)	161,334 (1)	283,572 (1837)	56,468 (402)	20,633 (1074)	37,222 (1869)	71.3	73.3	65.3
Rosa	60.5		134,461,068 (4)	41,929,383 (6)	812,407 (2187)	53,708 (672)	74397 (1866)	105,639 (1605)	88.8	97	97.2

Species	***Arabidopsis***	***Funaria***	***Liriodendron***	***Populus***	***Rosa***
RM%	23.6	42.35	73.18	35.9	60.53
RM2+%	16.5	43.1	72.7	45.1	
Total short-reads (total libraries)	511,277,126 (9)	549,205,030 (9)	1,408,831,670 (20)	267,403,772 (5)	134,461,068 (4)
Total long-reads (total libraries)	23,134,068 (4)		10,437,029 (1)	161,334 (1)	41,929,383 (6)
Total Trinity transcripts before aligning to genome (N50)	319,434 (2726)	151,265 (1198)	2,839,867 (3055)	283,572 (1837)	812,407 (2187)
Total Trinity transcripts after aligning to genome (N50)	47,024 (1410)	55,085 (1383)	152,255 (1335)	78,774 (1320)	95,016 (1272)
Total ST2 transcripts (SR) (N50)	37,747 (2538)	59,741 (369)	62,341 (1041)	56,468 (402)	53,708 (672)

Total ST2 transcripts (LR) (N50)		36,241 (1599)		33,895 (1815)	20,633 (1074)	74397 (1866)
Total ST2 transcripts (SR/LR) (N50)		42,265 (1363)		45,785 (2361)	37,222 (1869)	105,639 (1605)
BUSCO (%)	**Trinity (SR)**	82.7	72.9	92.5	71.3	88.8
	ST2 (SR)	95.5	84.5	87.1	73.3	97
	ST2 (SR/LR)	93.6		77.3	65.3	97.2

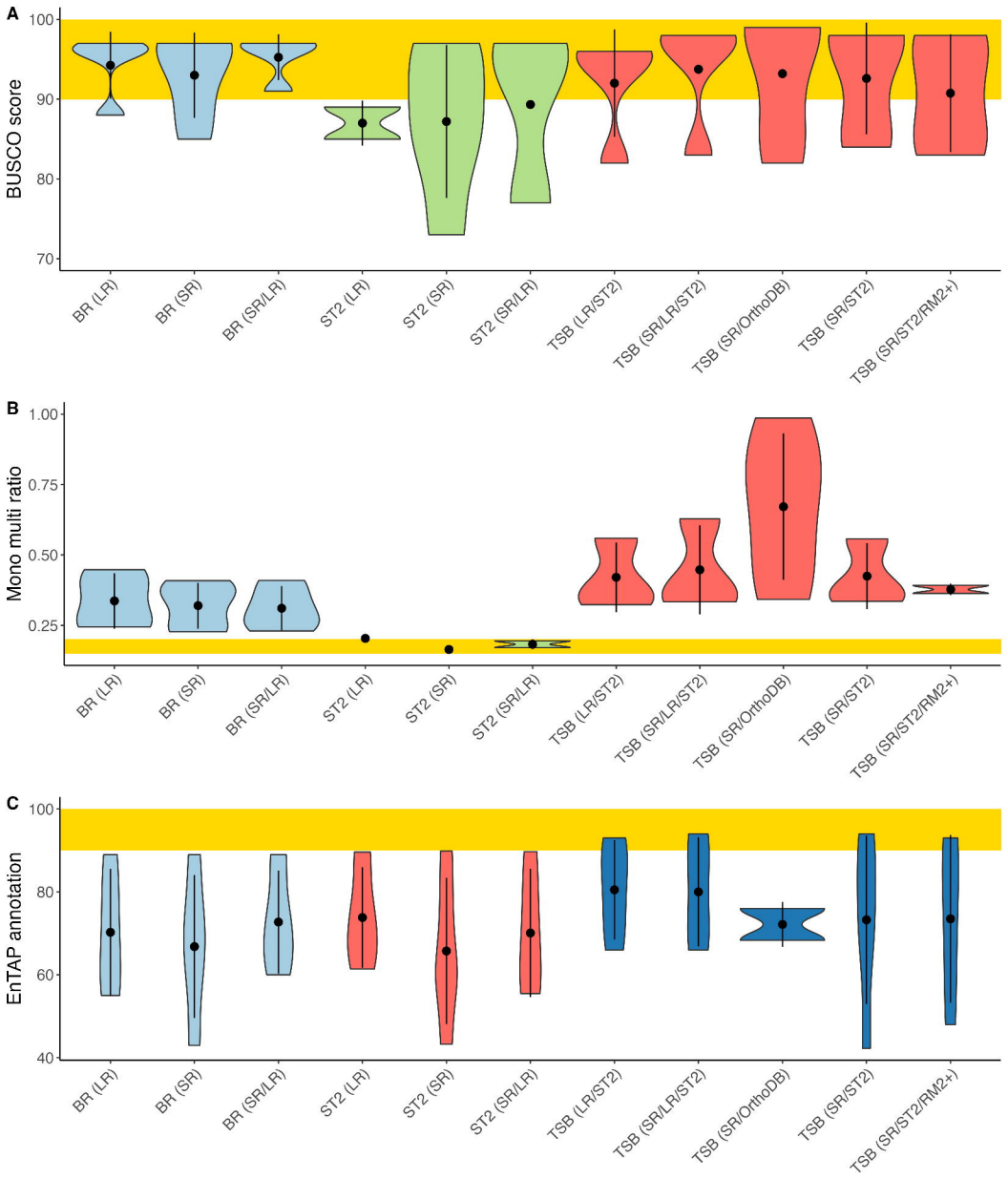
SR denotes short reads, LR for Long reads, RM for RepeatModeler2, and RM2+ for RepeatModeler2 with the LTRStruct flag, RM% indicates the repeat content masked with RepeatModeler2, and RM2+% is the repeat content masked with RepeatModeler2 with the LTRStruct flag.

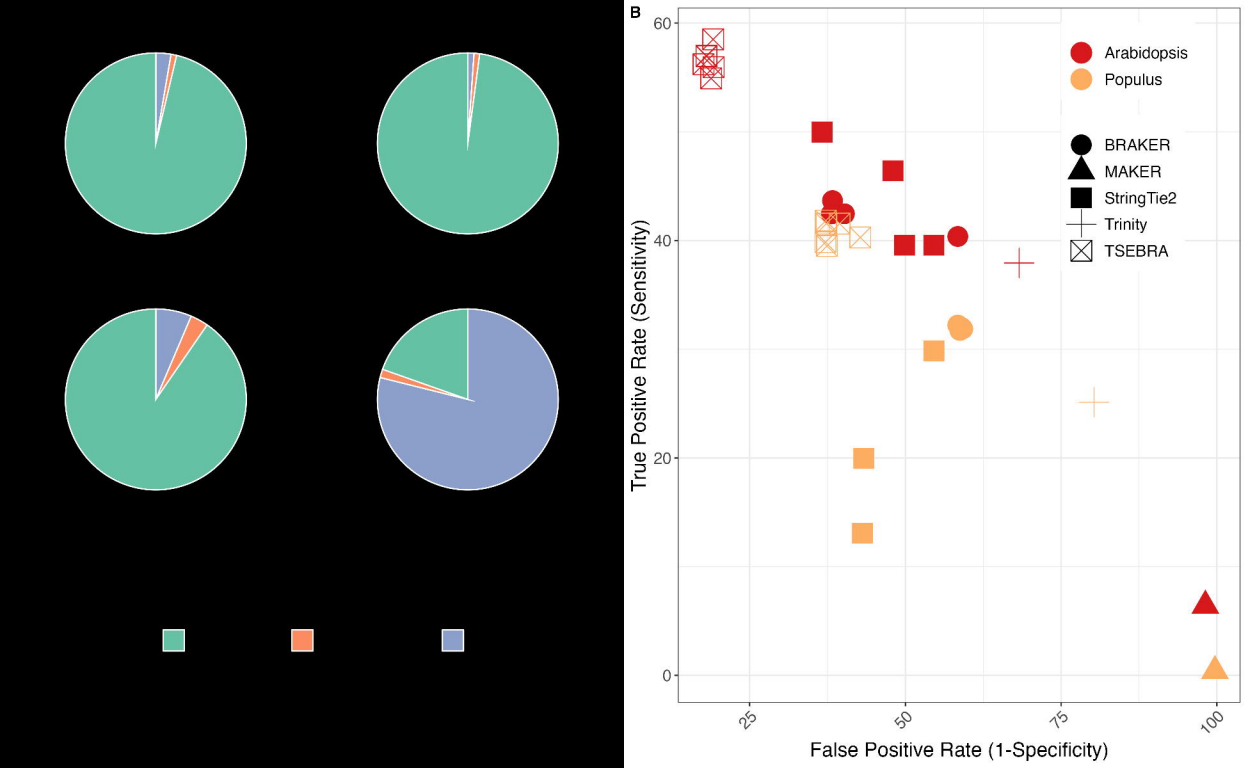
Table 4: Gene model statistics for Liriodendron after two rounds of structural and functional filters.

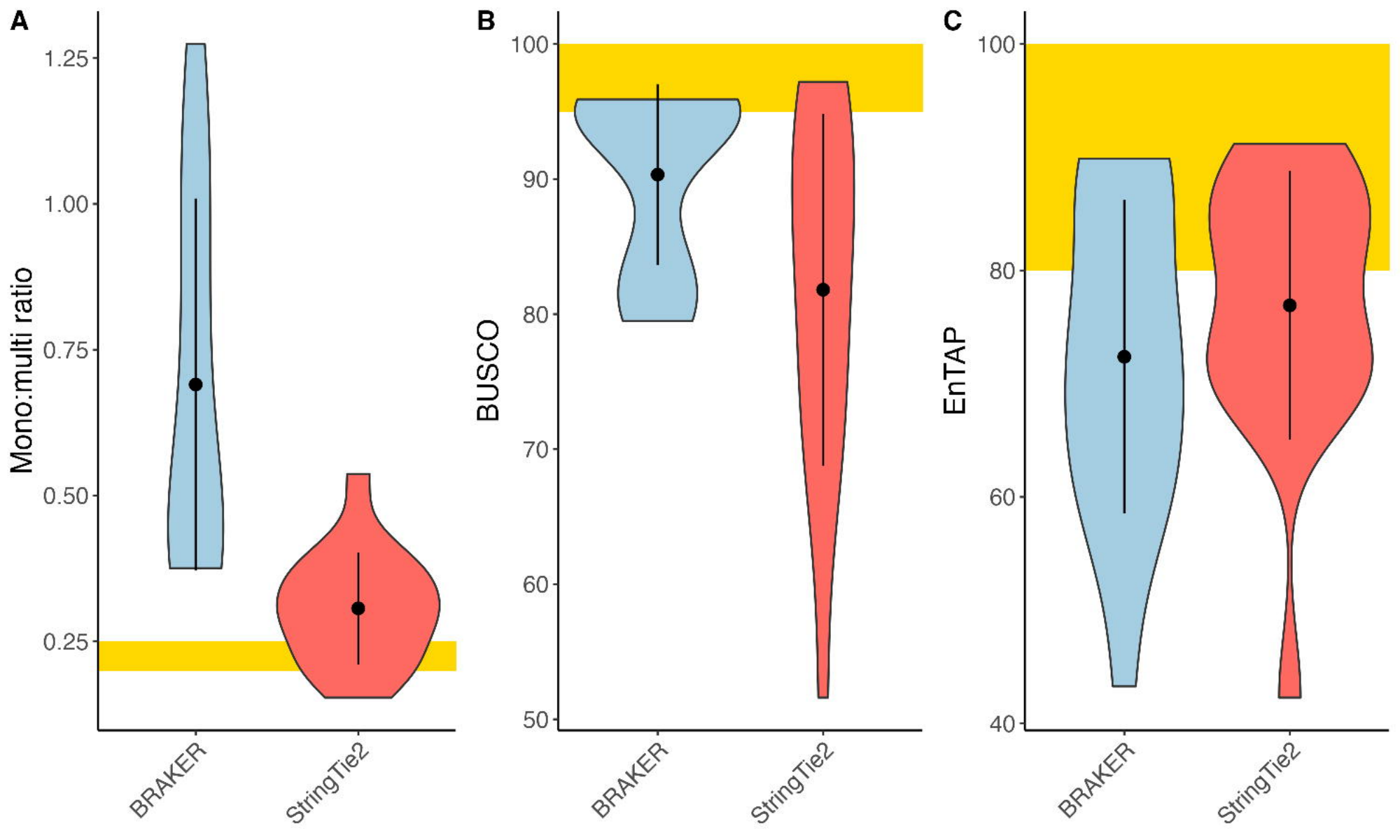
***Liriodendron* Annotation**	**Total genes**	**Mono:Multi Ratio**	**BUSCO %**	**EnTAP %**
Published annotation	35261	0.7	75.1	63
Mono-exonic filters				
BR (LR)	39031	0.21	87.4	69
BR (SR) *	41065	0.16	90.2	66
BR (SR/LR) *	40420	0.16	90.3	67
BR (SR/RM2+)	40740	0.17	88.2	67
ST2 (SR)	51804	0.16	86.5	80
ST2 (LR)	27012	0.23	65	84
ST2 (SR/LR)	36345	0.24	70.6	82
TSB (LR/ST2)	33132	0.43	82.3	84
TSB (SR/LR/ST2)	33964	0.43	82.4	84
TSB (SR/ST)	32898	0.41	83.4	84
TSB (SR/ST2/RM2+)	33637	0.45	82.8	84
TSB (SR/TRINITY)	34646	0.42	84	83
TSB (SR/OrthoDB)	33667	0.43	86.5	86
+Multi-exonic filters				
BR (SR)	30219	0.24	90.3	81
BR (SR/LR)	30035	0.23	86.9	87

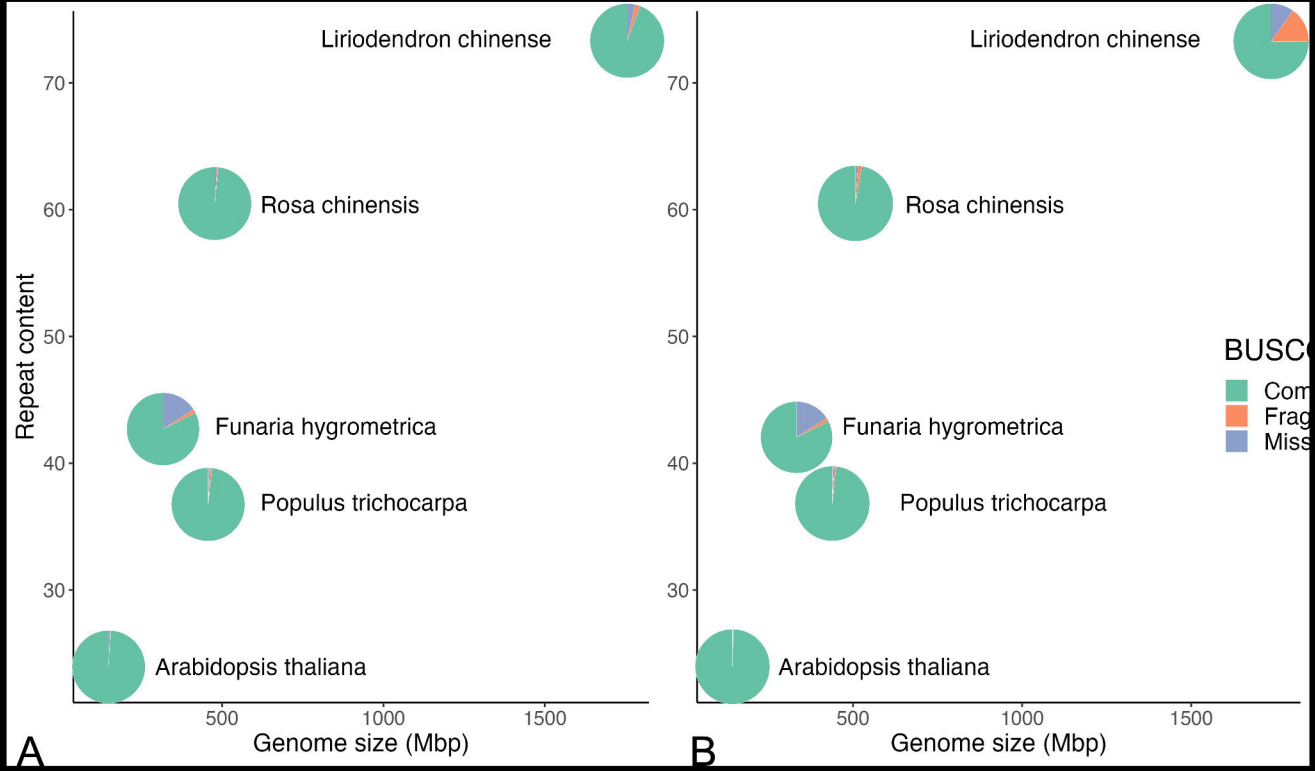
** denotes the two best annotation sets*

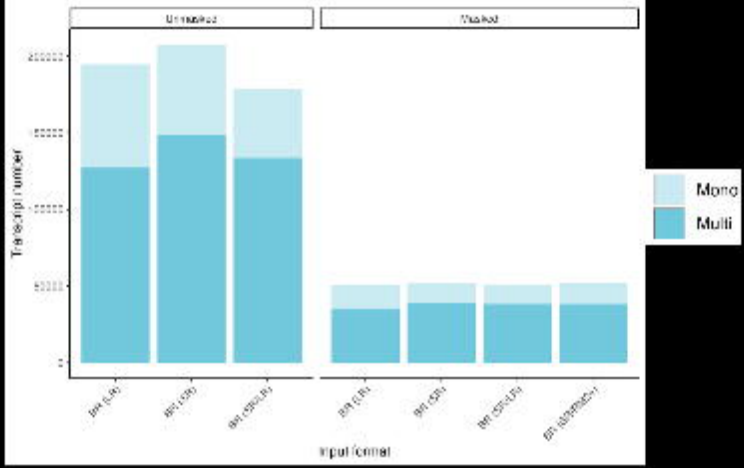
BR- BRAKER, ST2- StringTie2, TSB- TSEBRA, SR- Short reads, LR- Long-reads, RM2+- RepeatModeler2 with LTRStruct. BUSCO % shows the completeness percentage, and EnTAP % is the annotation rates after reciprocal BLAST.

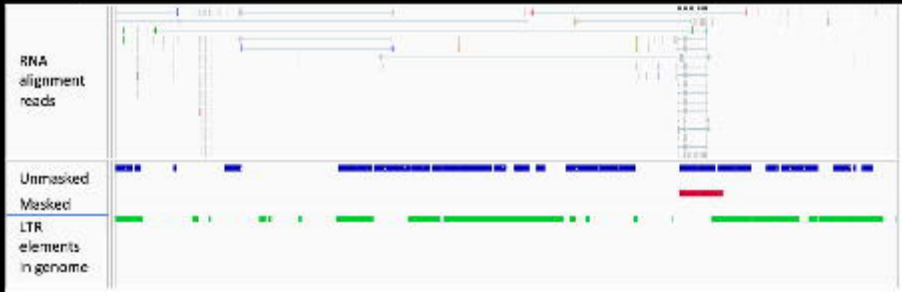

A
BUSCO score
100
90
80
70
BR (LR)
BR (SR)
BR (SR/LR)
ST2 (LR)
ST2 (SR)
ST2 (SR/LR)
TSB (LR/ST2)
TSB (SR/LR/ST2)
TSB (SR/OrthoDB)
TSB (SR/ST2)
TSB (SR/ST2/RM2+)
B
Mono multi ratio
1.00
0.75
0.50
0.25
BR (LR)
BR (SR)
BR (SR/LR)
ST2 (LR)
ST2 (SR)
ST2 (SR/LR)
TSB (LR/ST2)
TSB (SR/LR/ST2)
TSB (SR/OrthoDB)
TSB (SR/ST2)
TSB (SR/ST2/RM2+)
C
EnTAP annotation
100
80
60
40
BR (LR)
BR (SR)
BR (SR/LR)
ST2 (LR)
ST2 (SR)
ST2 (SR/LR)
TSB (LR/ST2)
TSB (SR/LR/ST2)
TSB (SR/OrthoDB)
TSB (SR/ST2)
TSB (SR/ST2/RM2+)
Software
BRAKER
StringTie2
TSEBRA


B
Arabidopsis
Populus
BRAKER
MAKER
StringTie2
Trinity
TSEBRA
True Positive Rate (Sensitivity)
False Positive Rate (1-Specificity)
0
20
40
60
25
50
75
100


A
Mono:multi ratio
1.25
1.00
0.75
0.50
0.25
BRAKER
StringTie2
B
BUSCO
100
90
80
70
60
50
BRAKER
StringTie2
C
EnTAP
100
80
60
40
BRAKER
StringTie2


Liriodendron chinense
Rosa chinensis
Funaria hygrometrica
Populus trichocarpa
Arabidopsis thaliana
Repeat content
70
60
50
40
30
500
1000
1500
Genome size (Mbp)
A
B
BUSC
Com
Frag
Miss


Unmasked
Masked
Transcript number
Mono
Multi
Input format


RNA alignment reads
Unmasked
Masked
LTR elements in genome