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* or genome-guided approaches 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, 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 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).

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

METHODS

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-

RepeatModeler2 (Flynn et al., 2020) was used to construct repeat libraries with default settings, and repeats were soft-masked with the libraries constructed via RepeatMasker v.4.0.6 (Smit et al. 2013-2015). The genomes of *Arabidopsis, Funaria, Populus, and Liriodendron* were additionally masked using RepeatModeler2 with additional LTR identification (-LTRStruct flag). Quality assessment of the Illumina short-reads was performed using FastQC v.0.11.7 (Andrews, 2010) before and after trimming low-quality bases. Sickle v.1.33 (Joshi NA, 2011) was used to trim low-quality bases with 50bp as the minimum read length threshold. Single-end reads generated post trimming were excluded from RNA alignments and assembly. The trimmed short reads were aligned against their reference genomes using HISAT2 v2.2.0 (Kim et al., 2019). HISAT2 was selected for its performance in recent benchmarking studies and as the aligner of choice for input to Stringtie2 (Corchete et al., 2020; Musich et al., 2021). Long-read RNA data were obtained for four species: *Arabidopsis* and *Rosa* were sequenced with Oxford Nanopore, and *Populus* and *Liriodendron* were sequenced with PacBio Sequel. The long-read data sets were aligned against their respective genomes using Minimap2 v2.1.7 (Li, 2018, 2021).

Generation of protein evidence-

To generate protein evidence, Illumina short reads were assembled *de novo* using Trinity v.2.8.5 with a minimum contig length of 300 bp (Grabherr et al., 2011). The assembled transcriptomes for the multiple libraries were combined, and putative coding regions were predicted using TransDecoder v.5.3.0 (http://transdecoder.github.io). TransDecoder is one of several frame-selection methods available and performs in a comparable manner but not always superior in all metrics (Bolger et al., 2018). For this study, it was selected as the most widely used package for this purpose. Redundancy in the predicted coding regions was reduced after clustering at 98% identity using UCLUST, a clustering algorithm of USEARCH v.9.0.2132 (Edgar, 2010). Frame-selected transcripts shorter than 300 bp were removed. The remaining 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.

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.

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. This annotation rate depends on the phylogenetic placement of species relative to the databases used, but the higher the annotation rate (>80%), the better the gene prediction assessment. The same is true for BUSCO since it utilizes OrthoDB to form its conserved sets but the recommended target score is >95% for land plants (Manni et al., 2021). 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 percentage statistics. The gene models predicted for *Liriodendron* were further filtered down and the mono-exonic and multi-exonic genes were filtered for uniqueness (using the unique genes flag in gFACs). The mono-exonic genes were filtered for the presence of protein domains using InterProScan v.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

Genome sizes, repeats, and published annotations-

The genome sizes of the five species assessed represented a 10-fold difference between the smallest genome of *Arabidopsis* (~119 Mb) and the largest of *Liriodendron* (~1.7 Gb) (Fig 1A; Table 2). *Liriodendron* (73.18%) and *Rosa* (60.58%) have higher levels of repeat content, and *Arabidopsis* has the lowest (23.9%). *Arabidopsis* is the most complete chromosome-scale genome, with seven contigs reflecting its five chromosomes and two organellar chromosomes. The other genomes are assembled into pseudochromosomes (with the exception of *Liriodendron*). Once the genomes were downloaded, contigs < 500 bp were removed. The published genome assemblies and annotations were compared in terms of completeness via BUSCO (Fig 1, Table 2). When BUSCO is run in genome mode, it searches the genome for the set of 1614 single-copy orthologs in the embryophyte database. Aside from *Funaria*, which had the lowest completeness score of 82.4%, the remaining plant genomes ranged from 94% to 99%. When we evaluated the published annotations for the same species, and ran BUSCO in protein mode, a slight decrease in completeness was observed in every species except *Funaria* and *Arabidopsis* (Fig 1B). The largest reduction in BUSCO score was observed in *Liriodendron* (98.6% to 75.1%). The discrepancy between the estimated completeness at the genome-level and the majority of the published annotations speaks to the challenges of achieving an accurate structural annotation.

RepeatModeler2 (RM2) with and without the LTRStruct package (the additional LTR masking module) (Flynn et al. 2020) was used to soft-mask repeats in four of the genomes. The increase in repeat content was marginal in all species, ranging from 1% in *Funaria* to 5% in *Populus*. Comparisons using the LTRStruct flag were denoted as RM2+ (Table S4).

Transcriptome evidence-

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

Transcript-derived annotations-

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

Arabidopsis and *Populus* were further evaluated with Mikado to compare the sensitivity and specificity of published annotations (Fig.4B, Table S7). Overall, StringTie2 predictions had higher sensitivity and precision rates compared to the Trinity runs. From this point, Trinity was excluded, and StringTie2 runs were compared against BRAKER and TSEBRA predictions (Fig 4B).

The mono:multi ratios produced by StringTie2 ranged from 0.15 in *Populus* (ST2 (LR)) to 0.53 in *Liriodendron* (ST2 (LR)), which were an improvement over the mono:multi ratios produced from the BRAKER annotations that ranged from 0.37 in *Arabidopsis* (BR (LR)) to 1.27 in *Funaria* (BR (SR/RM2+)). The BUSCO scores of the proteins predicted from BRAKER were generally higher than the BUSCO scores from StringTie2. For example, *Arabidopsis* StringTie2 runs range from 85% (ST2 (LR)) to 95.5% in ST2 (SR), and BRAKER runs ranged from 94% (BR (LR)) to 95.9% (BR (SR)). However, some runs are comparable, the ST2 (SR) run with a BUSCO score of 95% was similar to the BR (SR) run at 95% and the BR (SR/RM2+) run at 95% in *Arabidopsis*. StringTie2 predicted models had a higher annotation rate, in general, compared to BRAKER. For example, the EnTAP annotation rate in *Funaria* was just over 40% post BRAKER, but was near 60% from the StringTie2 runs (Fig 2).

Genome annotation with MAKER and BRAKER-

To replicate MAKER-P's repeat pipeline, the RM2+ genome was used for *Arabidopsis, Populus*, and *Funaria* for the MAKER runs. BUSCO completeness was low, compared with BRAKER runs, and ranged from 19.6% in *Populus* to 90.4% in *Arabidopsis* (Fig 3A). The mono:multi ratio of MAKER(RM2+) for *Arabidopsis* was comparable to the BRAKER runs for the same species (0.22 for BR (SR) and BR (SR/RM2+), 0.24 for BR (LR), and 0.23 for BR (SR/LR)). The MK (RM2+) predictions for the total number of genes in *Arabidopsis* and *Funaria* were in the expected range for these species, 22K and 44K genes, respectively; whereas, only 7K genes were predicted for *Populus*. The gene lengths ranged from 1.8 Kb in *Funaria* to 2.3 Kb in *Arabidopsis* (Table S8). The best run for MK (RM2+) was for *Arabidopsis*, with a mono:multi ratio of 0.22 and a BUSCO score of 90.4%. On the other hand, the mono:multi ratio for *Populus* was 0.07, and the BUSCO score was 19.6%.

The model systems, *Arabidopsis* and *Populus*, further were evaluated with Mikado to compare the sensitivity and specificity of the published annotations (Fig. 3B; Table S7). The sensitivity and precision scores for gene predictions were the lowest from MAKER then Trinity, and highest from TSEBRA runs. StringTie2 and BRAKER yielded similar sensitivity and specificity scores for *Arabidopsis*, whereas for *Populus* the sensitivity score was lower than those from BRAKER runs. Given its overall low scores, MAKER was excluded from the subsequent comparisons. It should be noted, however, the outcomes of MAKER can be improved through the inclusion of external programs, such as GeneMark-ES, from BRAKER (Brůna et al., 2021).

In general, BUSCO scores were higher in the BRAKER and TSEBRA runs compared to StringTie2 runs, mono:multi ratios were the lowest in the StringTie2 runs, and all methods performed equally in terms of annotation percentage (Fig 4). Overall, the gene models generated by BRAKER for *Arabidopsis* performed similarly according to BUSCO completeness scores. The mono:multi ratios across BRAKER runs ranged between 0.23 to 0.39, and the annotation percentage was consistently above 95%. Compared to the StringTie2 annotations, the BRAKER and TSEBRA runs had worse mono:multi ratio, and overall fewer genes. *Funaria* had more variable results in terms of mono:multi ratio, from 0.39 for BR (SR/RM2+), and 1.27

TSB (SR/ST2/RM2+). The annotation percentages for *Funaria* were lower than expected, 43% for BR (SR), and TSB (SR/TRINITY) had the highest annotation percentage with 53%. The BUSCO completeness scores of about 85% post BRAKER are comparable to those from StringTie2. In the case of *Liriodendron* post BRAKER, there were more variable mono:multi ratios as compared to the respective StringTie2 runs, which ranged from 0.34 to SR, and 1.04 BR (SR/RM2+). The annotation percentages for each run were around 75%, with BUSCO scores between 83% for TSB (SR/LR/ST2), and 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 percentages also differed between TSEBRA and BRAKER from 75% to 87%, respectively. *Rosa* had overall consistent scores for BUSCO post BRAKER, ranging around 96%. TSEBRA runs had higher mono:multi ratios of around 0.75 and 0.37 for BRAKER runs (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, with the exception of *Populus*. The average gene length post BR (LR) for *Populus* ranges from 2.7K to 3.4K, although some transcripts exceed 6 Kb in length. The longest predicted gene length was for a *Liriodendron* gene, estimated to be 9.3 Kb. The inclusion of long-reads (only) did not improve BUSCO completeness for any species, with the exception of *Arabidopsis*, where the BR (LR) BUSCO completeness was 1% higher than the BR (SR) run. The rise in BUSCO completeness in *Arabidopsis* could be due to the large number of long-reads included (23M across four libraries). However, the quality of genome annotation does not seem correlated with depth of long-read sequencing; for example, *Rosa* had more reads (41M across 6 libraries), and the BR (LR) run had a similar BUSCO score to BR (SR) (96%). It should be noted that the long-reads for *Arabidopsis* and *Rosa* were sequenced with

ONT. The ONT reads had higher mapping rates, compared to Iso-Seq, to their respective genomes, 97.1% in *Arabidopsis* and 99% in *Rosa* (Table S5). The long-read inputs, regardless of depth or type, impact ST2 (LR) runs across all species, with a reduction of up to 10% in BUSCO completeness (Table S9). Finally, we note that the combination of short-reads and long-reads BR (SR/LR) is comparable to the BR (SR) reads in terms of BUSCO completeness, annotation rate and total genes predicted, but had worse mono:multi ratios overall.

Refining the genome annotation for Liriodendron-
The BRAKER runs for *Liriodendron* were filtered with gFACs and InterProScan to remove unlikely gene models (Table 4). The number of mono-exonic genes was drastically reduced post-filter with InterProScan. Across all runs, the mono-exonic genes numbered 11K to 25K. After removing mono-exonics without a protein domain annotation from the Pfam database, they decreased from 11K to 5K. The decrease in false positive mono-exonics resulted in an improved mono:multi ratio that nor range from 0.16 for BR (SR) and BR (SR/RM2+), 0.16 and 0.23 for the StringTie2 runs, to 0.43 for the TSEBRA runs. The BUSCO scores decreased slightly post-filtering (1-2%). EnTAP annotation percentages ranged between 66% to 84%, with the TSEBRA runs, and ST2(LR) having the highest annotation rates overall.

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

DISCUSSION

BRAKER (Hoff et al., 2020) and MAKER (Cantarel et al., 2008) are currently the most popular eukaryotic structural annotation tools, cited 475 and 1,010 times (since 2021, as referenced in Google Scholar). Processes that select from multiple *ab initio* or aligned forms of evidence are gaining popularity as well though they add both time and complexity to the analyses (FINDER cited 22 times, (Banerjee et al., 2021); EVidenceModeler cited 381 times (Haas et al., 2008)). Finally, as high-throughput transcriptomics, in the form of both short and long-read evidence become more accessible, rapid approaches like StringTie2 (cited 451 times (Kovaka et al., 2019)) are occasionally used as the exclusive approach, though more often, used in combination with the options listed above.

Regardless of the methods selected, recently published benchmarks are challenged to achieve high values for gene sensitivity in larger genomes (Brůna et al., 2021). Within smaller and less 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 17 BRAKER runs and four of the 17 StringTie2 runs exceeded 95% completeness. However, total gene number, gene length, and structure varied considerably.

Repeat content, especially in the form of LTRs, and pseudogenes can lead to inflated gene model estimates, especially in the form of mono-exonic genes (Scott et al., 2020; Trouern-Trend et al., 2020). We expect that eukaryotes maintain 20% or less of their gene space as mono-exonics (Jain et al., 2008). Although the BUSCO scores were consistent, we note tremendous variation in mono- to multi-gene model ratios post-BRAKER. In practice, having a worse mono:multi ratio is preferable to having a lower BUSCO score, since missing genes, especially those thought to be conserved, cannot be easily rectified, and putative false positives can potentially be filtered through other means.

Sequence similarity search metrics are more complex to interpret, but when used with high-quality and curated databases that contain full-length proteins (e.g. NCBI RefSeq), can provide a benchmark. Specifically, a reciprocal BLAST search requires that both the query and target in the search retain a minimum level of coverage in the alignment. For new plant genomes, that are in the darkest corners of the tree of life, this might be a less reliable metric. For species that may fare poorly in database comparisons, searches for protein domains can provide some level of confidence and we demonstrate this as a filter to reduce the mono-exonics in *Liriodendron.*

Masking repeats in plant genomes: Repeat masking is important but may not require additional LTR resolution to improve performance-

Plant genomes typically contain a large number of repeats, mostly in the form of transposable elements (TEs), averaging around 46% (Luo et al., 2022). Given the abundance of TEs in genomes, it is important to mask these in advance of gene prediction. Soft-masking involves changing nucleotides identified as repeats to lowercase (Yandell & Ence, 2012), signaling downstream programs to ignore these sequences. Of the five genomes included in this study, *Liriodendron* had the largest genome size and repeat content. Running downstream analyses on an unmasked genome of *Liriodendron* resulted in a 4-fold increase in gene predictions (Fig 5A). Many repeats were identified as putative gene models, resulting in a large increase of total number of genes (Fig 5B).

RepeatModeler2 is a widely used tool for TE discovery (Flynn et al., 2020). The recent release of RepeatModeler2 includes an optional module for more robust LTR structural detection (LTRStruct module) that includes the LTRharvest (Ellinghaus et al., 2008), LTRDetector (Valencia & Girgis, 2019), and LTR_retriever packages (Ou & Jiang, 2018). This is particularly useful in identifying more divergent LTRs in the genome that may exist in fewer copies (Ou & Jiang, 2018; Valencia & Girgis, 2019). Among the default packages included, RepeatScout serves as a fast method to detect young and abundant repeat families in the genome. RECON, on the other hand, is more computationally intensive and is sensitive enough to detect older TE families. The LTRStruct module is run on the unmasked genome to identify LTR families that may be redundant with the families identified by the default package. This creates redundancy that is resolved through clustering with CD-HIT (Flynn et al., 2020).

In the four species compared, additional repeat masking did not significantly improve gene predictions (Table S9; Fig 4). The mono:multi ratios across species were consistent before and after additional LTR masking (Fig 5A). The BUSCO completeness scores remained relatively the same, with BR (SR/RM2+) being 1% higher than BR (SR) in *Arabidopsis, Funaria* and *Populus.* The marginal improvement observed in these genomes could be related to the structure and type of LTRs, for example, better identification of divergent Ty1-copia elements described in the *Funaria* genome (Kirbis et al., 2022). While we did not include genomes with excessive repeat estimates (>70%), our results indicated that the optional LTRStruct module was not beneficial.

Genome-guided transcriptome assembly for annotation: Transcripts derived directly from alignments are not sufficient to annotate reference genomes-

Transcriptome assemblers are designed to work with primarily short RNA-Seq reads to construct full-length transcripts. In the presence of a high quality reference genome, genome-guided approaches are preferred as the reads are aligned directly to the target genome in advance. Aligned RNA evidence provides resolution on exon boundaries, and aids in the identification of splice variants. *De novo* approaches build graph models directly from the short (or long) reads to generate transcripts. The latter is much more challenging, computationally intensive, and prone to error.

We compared the accuracy of the annotations produced by StringTie2, *de novo* assembled transcripts with Trinity and with BRAKER. The selected packages are top performers when compared in their respective categories of genome-guided and *de novo* transcriptome assembly (Sahraeian et al., 2017; Venturini et al., 2018). As expected, Trinity produced a higher number of transcripts than StringTie2, and BUSCO completeness was consistently lower (Table 3), except for *Liriodendron*. The gene models generated by StringTie2 were more numerous than the BRAKER gene models, more than expected for each species. It should be noted, however, that StringTie2 identifies splice variants by generating a splice graph and resolving conflict between multiple potential splice sites (Kovaka et al., 2019), whereas BRAKER trains an internal algorithm GeneMark-ET to find specific genes with complete support among all introns to be further used in training Augustus (Hoff et al., 2019).

StringTie2 runs resulted in lower BUSCO completeness when compared to BRAKER and/or TSEBRA runs (Fig 4A; Table S11). This outcome is supported by the lack of *ab initio* prediction with genome-guided approaches. Inflated mono-exonic predictions (and lower BUSCO scores) were also observed in the StringTie2 genome annotation of the water strider (*Microvelia longipes*) (Toubiana et al., 2021). In our study, *Rosa* ST2 (SR/LR) run was closest with a BUSCO score of 97.2%, BR (SR/LR) of 96.9%, and TSB (SR/ST2) of 98% (Table S9, S10).

Including proteins: Genome annotations are marginally improved if protein evidence sourced from genome-guided predictions is used in combination with read data for high quality reference genomes only-

The performance of StringTie2 and Trinity-derived protein evidence was assessed on the predicted gene models using BRAKER and TSEBRA. In this context, the genome-guided or *de novo* assembled transcripts were translated into proteins and provided as evidence to train the *ab initio* component of the pipelines. Adding protein evidence to genome annotation can target protein-coding genes leading to more accurate predictions than RNA-Seq evidence alone (Bruna, 2022). This study specifically focused on using protein evidence derived in some fashion from the transcriptomic inputs; however, some workflows, including BRAKER, recommend including external curated protein sets (i.e., UniProt, RefSeq) to provide additional evidence. We avoided this comparison since the protein models would represent the true protein models for the model systems, but it is worth noting that this could improve some outcomes.

The TSEBRA runs of the model species, *Arabidopsis* and *Populus* were compared to the reference annotations. These runs were the best for the model species in terms of sensitivity and specificity as compared to the MAKER, StringTie2, Trinity and BRAKER runs (Fig 3B). The model genomes also had very similar BUSCO completeness scores, but had different mono:multi ratios with the addition of protein evidence The non-model plant genomes had higher mono:multi ratios, and especially in the case of *Liriodendron*, the BUSCO scores were overall lower than the non-protein runs. The *Rosa* TSB (SR/ST2) reported the highest BUSCO score across all runs but at the expense of more putative false positives. The higher quality of the *Rosa* genome assembly could influence the utility of the protein evidence.

TSEBRA runs with proteins sourced from genome-guided predictions perform similarly, but had lower BUSCO, mono:multi ratio and total gene number when compared to the SR only runs (Fig 4A, Table S9). Among TSEBRA runs, Trinity fares better only for *Liriodendron*, which could indicate that genome-guided proteins are not a suitable choice for a more fragmented genome. This is consistent in independent assessments between *de novo* transcriptome assemblers and genome-guided assemblers with fragmented genomes (for example in *Ae. albopictus* (Huang et al., 2016). The total number of genes predicted by TSEBRA and BRAKER runs remained largely the same across all species (Table S9). However, the number of mono-exonic genes increased, whereas the multi-exonic genes decreased across all TSEBRA runs in comparison to the BRAKER runs without proteins across all species. The gene lengths also decreased, as expected from the increase in mono-exonics.

Initial examination of the EnTAP reciprocal BLAST assessment revealed high annotation rates for the non-model species when protein evidence was included, particularly the multi-exonics (whereas the mono-exonic percentage remained the same) (Table S9). However, this increase in multi-exonic annotation proved to be an artifact since the total number of multi-exonic models was reduced. Direct comparison of the predictions revealed that 40% of multi-exonics were actually split into mono-exonic predictions when comparing the BR(SR) to the TSB (SR/ST2) gene models predicted using *Liriodendron* (Table S15).

Long-read transcriptomes: Long-reads can be paired with short-reads to improve the quality of the resulting models-

Long-reads generated from platforms such as Oxford Nanopore or PacBio have the potential to resolve splice variants and assemble transcripts more accurately than traditional Illumina RNA-

Seq (Amarasinghe et al., 2020). While long-reads can independently generate transcriptomes, it is recommended to have a combination of short and long reads to achieve greater depth, improved error profiles, and gain more evidence for splice site resolution (Amarasinghe et al., 2020; Gonzalez-Ibeas et al., 2016; Watson & Warr, 2019).

In this study, we utilized both ONT and Iso-Seq long-reads. In the latter, we relied on raw reads (not the error-corrected CCS reads) in our comparisons for genome annotations using long-reads. In all cases, long-reads (alone) did not outperform short-reads for the BRAKER runs. However, in some cases, the combination of short-read and long-read inputs was beneficial. The higher error rate Iso-Seq reads from *Populus* and *Liriodendron* produced comparable, but lower, BUSCO scores compared to the BR (SR) runs. In contrast, the ONT long-reads used for *Arabidopsis* and *Rosa* in the combined runs (BR (SR/LR)) had slightly better BUSCO completeness as compared to the BR (SR) runs, and similar mono:multi ratios. Overall, the lower error profile of using ONT reads, supplemented with short-read data, as well as using high quality reference genomes, support the higher BUSCO completeness scores.

Best Practices for Plant Genome Annotation-

Given existing tools, we recommend that investigators utilize RepeatModeler2 to mask their genome of interest with the default settings available in v2 (Flynn et al. 2020). Following soft-masking, RNA-Seq short reads (between 4-10 libraries, paired-end, minimum 15M reads per library) are generally sufficient for annotation. While we did not comprehensively investigate the impact of tissue type, it is recommended to sample from multiple tissues when possible (Kress et al., 2022). In our study, we did not observe a difference in the annotation completeness among species with a higher number of short-read libraries, although we did not comprehensively evaluate the difference of using fewer libraries within a single species.

Sequencing of long-read libraries remains more expensive than generating deep Illumina short-read RNA-Seq. In most cases, the short-reads were sufficient as input. The notable exceptions include the BR (SR/LR), as they were comparable to only BR (SR) across most species. The lower error-profile Nanopore reads were more beneficial when combined with short-reads.

BRAKER and TSEBRA outperformed runs of MAKER, StringTie2 and Trinity with default settings. It should be noted that the authors did not comprehensively benchmark MAKER with multiple training runs of AUGUSTUS as recommended, which could have further improved results. However, previous benchmarking studies also support lower performance of MAKER (Banerjee et al., 2021; Hoff et al., 2020). Among the BRAKER runs executed in the model plants, *Arabidopsis* and *Populus*, the TSEBRA runs were the best runs. TSEBRA also appears to perform the best for *Rosa* but would require substantial filtering to remove false-positives. Among the less contiguous assemblies (*Funaria* and *Liriodendron*), BR (SR) runs performed the best in terms of BUSCO completeness, mono:multi ratios, and EnTAP annotation rates. For draft genomes, BRAKER runs with short-reads, or short-reads and long-reads, when high quality long-read transcripts from deeper sequencing are available, is advised.

Regardless of approach, existing pipelines do not provide appropriate summary statistics or robust methods for filtering unlikely gene models. All methods produce more putative false positives than desired. We recommend utilizing reciprocal BLAST searches with well curated databases containing targets with full-length proteins (such as NCBI’s RefSeq) to identify fragmented models. We also recommend filtering and removing mono-exonics that do not have a protein domain. Finally, we recommend structural filters to remove unlikely gene structures (splice sites, start sites, incompletes, etc).

In this study, we demonstrated the impact of post-filtering on the most complex genome assessed in this study, *Liriodendron*. We improved the published annotation across all 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.

LITERATURE CITED

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), 30.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. Available online. *Retrieved May*, *17*, 2018.

Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, *408*(6814), 796–815.

Banerjee, S., Bhandary, P., Woodhouse, M., Sen, T. Z., Wise, R. P., & Andorf, C. M. (2021). FINDER: an automated software package to annotate eukaryotic genes from RNA-Seq data and associated protein sequences. *BMC Bioinformatics*, *22*(1), 205.

Bolger, M. E., Arsova, B., & Usadel, B. (2018). Plant genome and transcriptome annotations: from misconceptions to simple solutions. *Briefings in Bioinformatics*, *19*(3), 437–449.

Bruna, T. (2022). *Unsupervised algorithms for automated gene prediction in novel eukaryotic genomes*. https://smartech.gatech.edu/handle/1853/67297

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.

Caballero, M., & Wegrzyn, J. (2019). gFACs: Gene Filtering, Analysis, and Conversion to Unify Genome Annotations Across Alignment and Gene Prediction Frameworks. *Genomics, Proteomics & Bioinformatics*, *17*(3), 305–310.

Campbell, M. S., Holt, C., Moore, B., & Yandell, M. (2014). Genome Annotation and Curation Using MAKER and MAKER-P. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, *48*, 4.11.1–39.

Campbell, M. S., Law, M., Holt, C., Stein, J. C., Moghe, G. D., Hufnagel, D. E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C. J., Ware, D., Shiu, S.-H., Childs, K. L., Sun, Y., Jiang, N., & Yandell, M. (2014). MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations. *Plant Physiology*, *164*(2), 513–524.

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., & Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, *18*(1), 188–196.

Chang, C., Bowman, J. L., & Meyerowitz, E. M. (2016). Field Guide to Plant Model Systems. *Cell*, *167*(2), 325–339.

Cheng, S., Melkonian, M., Smith, S. A., Brockington, S., Archibald, J. M., Delaux, P.-M., Li, F.-W., Melkonian, B., Mavrodiev, E. V., Sun, W., Fu, Y., Yang, H., Soltis, D. E., Graham, S. W., Soltis, P. S., Liu, X., Xu, X., & Wong, G. K.-S. (2018). 10KP: A phylodiverse genome sequencing plan. *GigaScience*, *7*(3), 1–9.

Chen, J., Hao, Z., Guang, X., Zhao, C., Wang, P., Xue, L., Zhu, Q., Yang, L., Sheng, Y., Zhou, Y., Xu, H., Xie, H., Long, X., Zhang, J., Wang, Z., Shi, M., Lu, Y., Liu, S., Guan, L., … Shi, J. (2019). Liriodendron genome sheds light on angiosperm phylogeny and species-pair differentiation. *Nature Plants*, *5*(1), 18–25.

Corchete, L. A., Rojas, E. A., Alonso-López, D., De Las Rivas, J., Gutiérrez, N. C., & Burguillo, F. J. (2020). Systematic comparison and assessment of RNA-seq procedures for gene expression quantitative analysis. *Scientific Reports*, *10*(1), 19737.

Deutekom, E. S., Vosseberg, J., van Dam, T. J. P., & Snel, B. (2019). Measuring the impact of gene prediction on gene loss estimates in Eukaryotes by quantifying falsely inferred absences. *PLoS Computational Biology*, *15*(8), e1007301.

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

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

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

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

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

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

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

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

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

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

Wegrzyn, J. L. (2020). EnTAP: Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. *Molecular Ecology Resources*, *20*(2), 591–604.

Hoff, K. J., Brŭna, T., Lomsadze, A., & Stanke, M. (2020). Fully Automated and Accurate Annotation of Eukaryotic Genomes with BRAKER2. *Poster Presented at*. https://www.researchgate.net/profile/Katharina-Hoff-2/publication/338831355_Fully_Automated_and_Accurate_Annotation_of_Eukaryotic_Genomes_with_BRAKER2/links/5e2d9102299bf152167f6424/Fully-Automated-and-Accurate-Annotation-of-Eukaryotic-Genomes-with-BRAKER2.pdf

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., & Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* , *32*(5), 767–769.

Hoff, K. J., Lomsadze, A., Borodovsky, M., & Stanke, M. (2019). Whole-Genome Annotation with BRAKER. *Methods in Molecular Biology* , *1962*, 65–95.

Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*, 491.

Huang, X., Chen, X.-G., & Armbruster, P. A. (2016). Comparative performance of transcriptome assembly methods for non-model organisms. *BMC Genomics*, *17*, 523.

Jain, M., Khurana, P., Tyagi, A. K., & Khurana, J. P. (2008). Genome-wide analysis of intronless genes in rice and Arabidopsis. *Functional & Integrative Genomics*, *8*(1), 69–78.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics* , *30*(9), 1236–1240.

Joshi NA, F. J. N. (2011). *Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files* (Version Version 1.33) [Computer software]. https://github.com/najoshi/sickle

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

Toubiana, W., Armisén, D., Dechaud, C., Arbore, R., & Khila, A. (2021). Impact of male trait exaggeration on sex-biased gene expression and genome architecture in a water strider. *BMC Biology*, *19*(1), 89.

Trouern-Trend, A. J., Falk, T., Zaman, S., Caballero, M., Neale, D. B., Langley, C. H., Dandekar, A. M., Stevens, K. A., & Wegrzyn, J. L. (2020). Comparative genomics of six Juglans species reveals disease-associated gene family contractions. *The Plant Journal: For Cell and Molecular Biology*, *102*(2), 410–423.

Valencia, J. D., & Girgis, H. Z. (2019). LtrDetector: A tool-suite for detecting long terminal repeat retrotransposons de-novo. *BMC Genomics*, *20*(1), 450.

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

Watson, M., & Warr, A. (2019). Errors in long-read assemblies can critically affect protein prediction [Review of *Errors in long-read assemblies can critically affect protein prediction*]. *Nature Biotechnology*, *37*(2), 124–126.

Wu, T. D., & Watanabe, C. K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* , *21*(9), 1859–1875.

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews. Genetics*, *13*(5), 329–342.

FIGURES

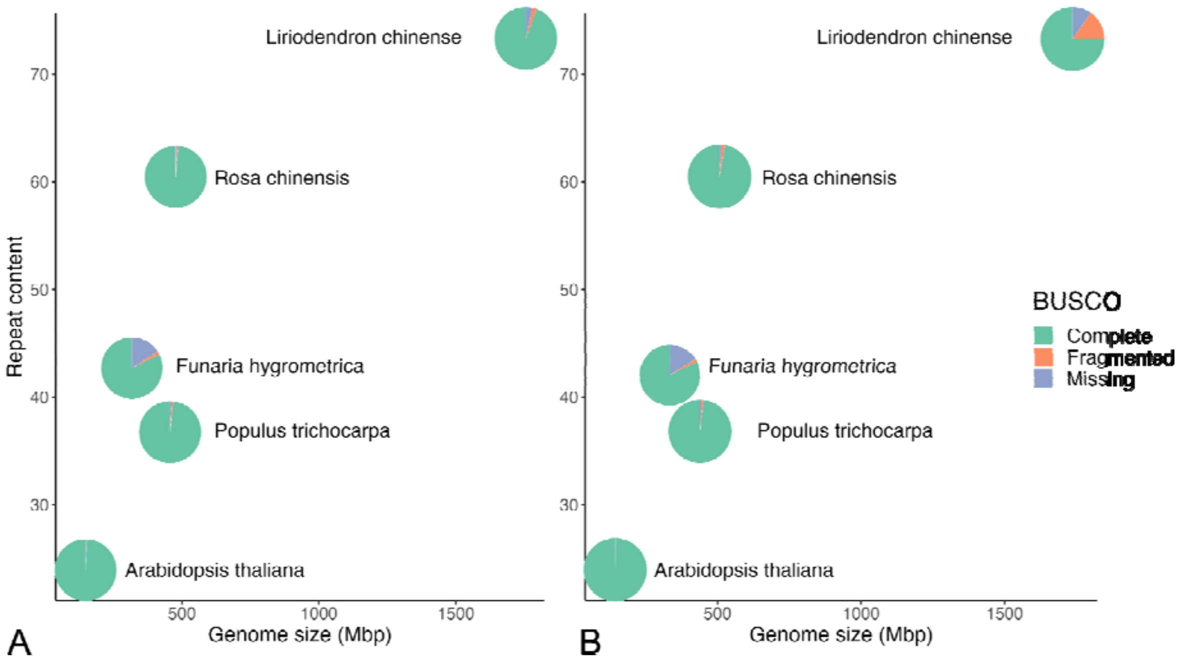

Figure 1: Genome size, repeat content, and BUSCO completeness for the five plant genomes: Arabidopsis, Populus, Funaria, Rosa, and Liriodendron. Each pie represents the BUSCO completeness. Green denotes the completeness score, orange indicates the fragmented score, and blue indicates the missing score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

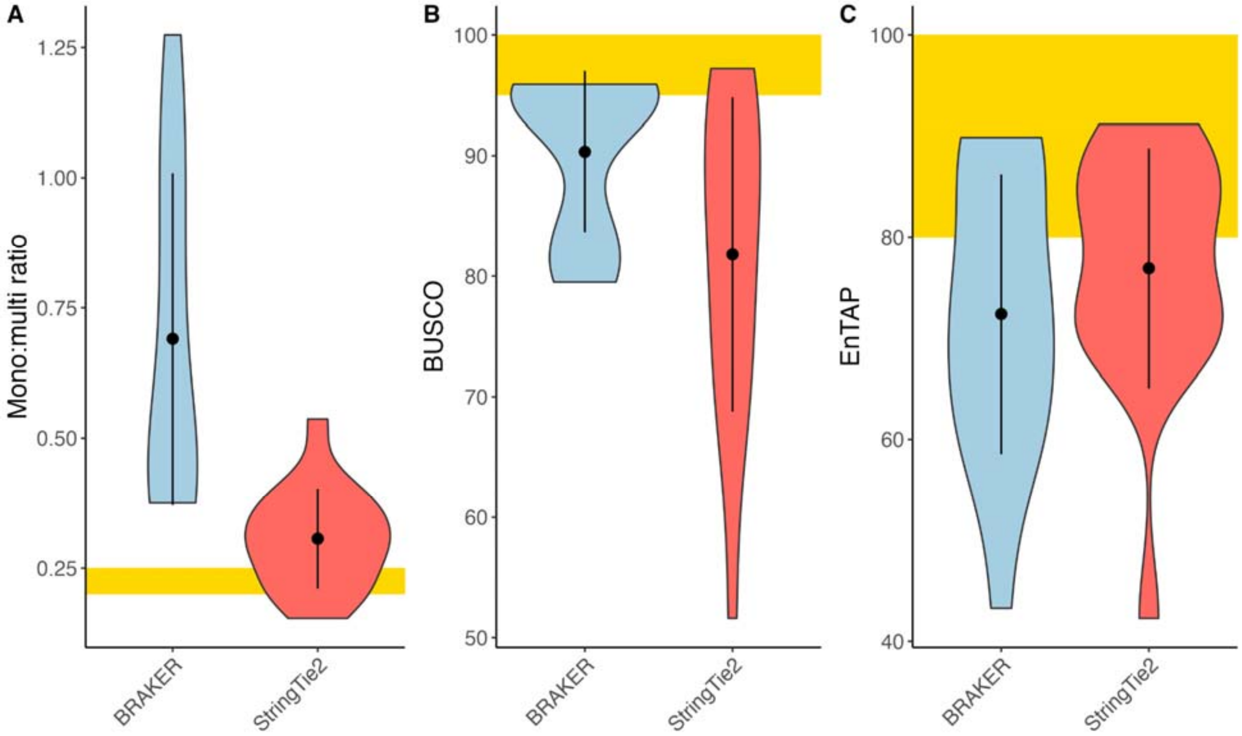

Figure 2: Comparing metrics between BRAKER (blue) and StringTie2 (red) predictions. (A) mono:multi ratios, (B) BUSCO comparisons, and (C) EnTAP annotation percentages of the gene models. The yellow region indicates the ideal value for each of the metrics.

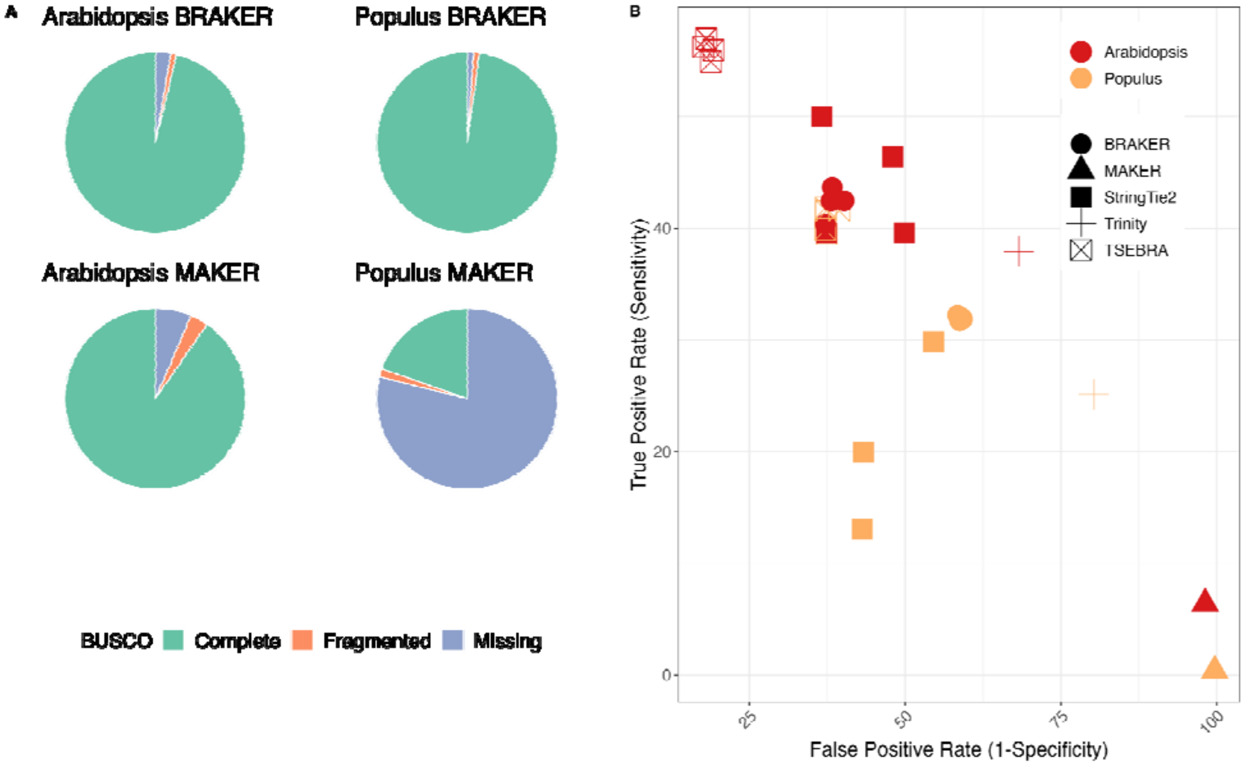

Figure 3: Comparison of BUSCO, sensitivity, and false positive rates between the Arabidopsis and Populus annotations. (A) BUSCO completeness scores for the MAKER and BRAKER runs of Arabidopsis and Populus, green denotes the completeness score, orange indicates the fragmented score, and blue indicates 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, Trinity, and StringTie2 runs. The scores were assessed using MIKADO. Multiple points per run reflect differences in input read type and repeat-masking.

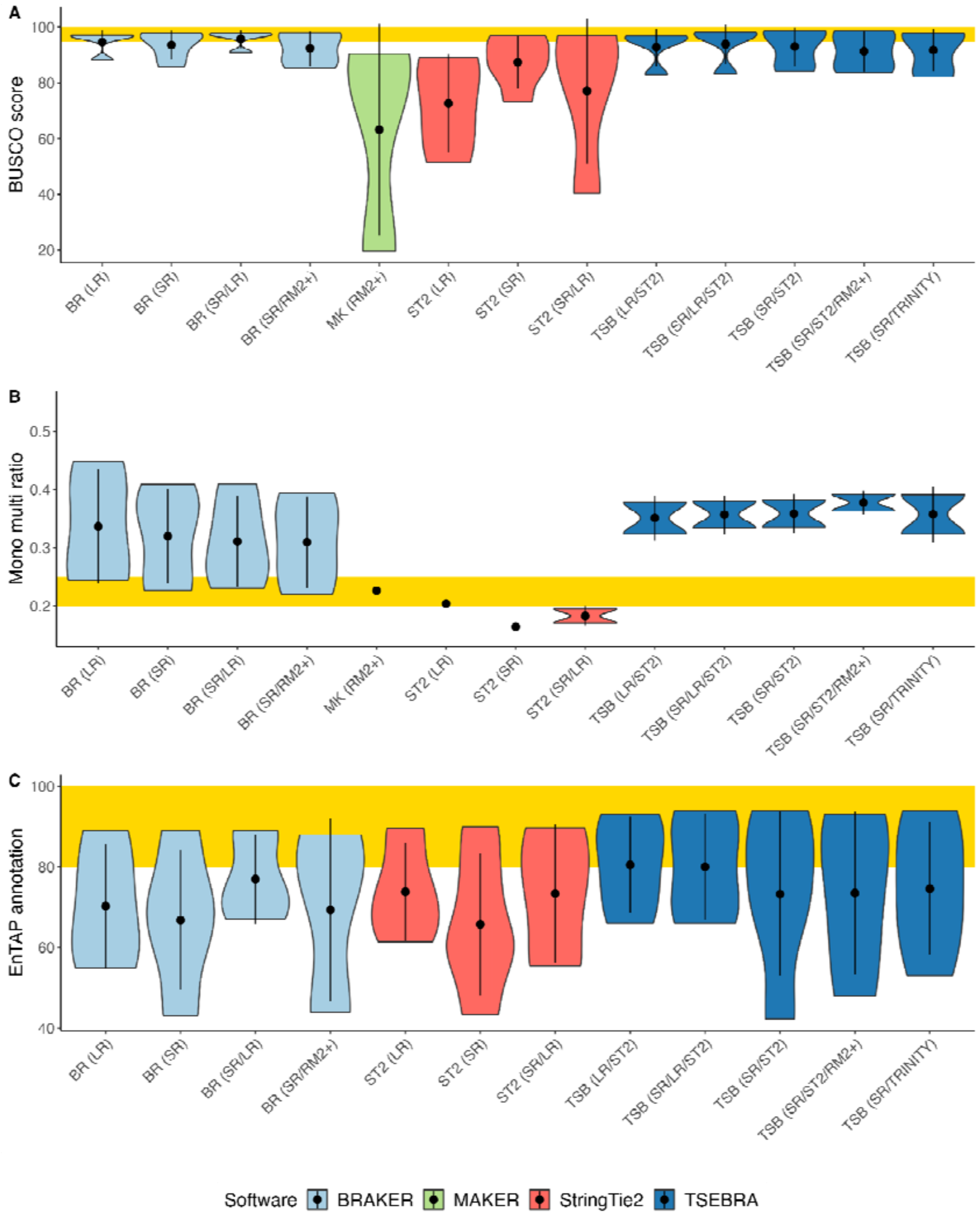

Figure 4: Comparison of mono:multi ratios (A), BUSCO completeness scores (B), and EnTAP annotation rates (C) across all species between the runs of different input types and software, i.e., MAKER (MK-green) BRAKER (BK-light blue),TSEBRA (TSB-dark blue) and StringTie2

(ST2-red). The yellow rectangle represents the target scores for each benchmark. RM2+-RepeatModeler2 with LTRStruct.

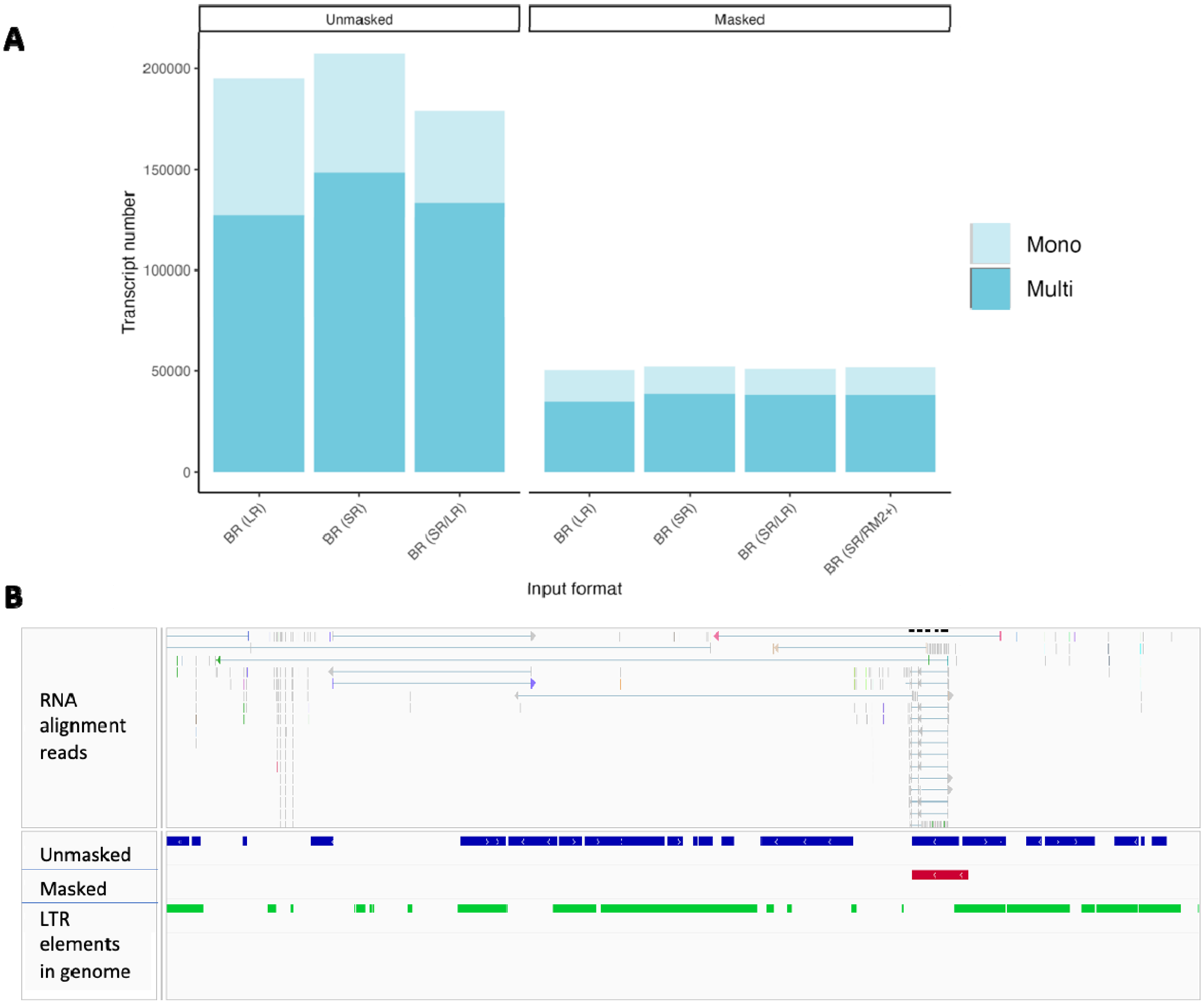

Figure 5: (A) The effect of soft-masking on gene prediction in Liriodendron. Performing structural annotation on the unmasked Liriodendron genome results in inflation in the mono and multi- exonic genes. Blue denotes the BRAKER (BR) runs for both genomes, SR denotes short-reads, and LR denotes long reads. The lighter shade represents mono-exonics, and the darker shade represents the multi-exonics. (B) More genes predicted using the unmasked genome (blue), as compared to only one gene predicted in this region with the masked genome (red).

The green track shows the LTR elements in the genome as identified by RepeatModeler2. The RNA alignment reads show a read pile-up at the predicted gene (masked track).

TABLES

Table 1: Notations for the different runs performed for benchmarking. SR- Short reads, LR- Long reads, and RM2+- RepeatModeler2 with the additional repeat masking.

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	
MAKER						
MK (RM2+)	Short-reads with additional masking for LTRs	X	X	X		

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

					BUSCO Completeness	
Species	**Genome size**	**Total scaffolds (chromosomes)**	**N50**	**Repeat content**	**Genome**	**Annotation**

Arabidopsis thaliana	119 Mb	7 (5)	23.46 Mb	15.2%	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. SR denotes short reads, LR for Long reads, RM for RepeatModeler2, and RM2+ for RepeatModeler2 with the LTRStruct flag.

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

Table 4: Gene model statistics for Liriodendron after two rounds of structural and functional filters. BR- BRAKER, ST2- StringTie2, TSB- TSEBRA, SR- Short reads, LR- Long-reads, RM2+- RepeatModeler2 with LTRStruct.

***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
+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*