FINDER: An automated software package to annotate eukaryotic genes from RNA-Seq data and associated protein sequences

Sagnik Banerjee[1,2], Priyanka Bhandary[1,3], Margaret Woodhouse[4], Taner Z. Sen[5], Roger P. Wise[4,6], Carson M. Andorf[4,7]*

[1] Program in Bioinformatics & Computational Biology, Iowa State University, Ames, IA 50011, USA

[2] Department of Statistics, Iowa State University, Ames, IA 50011, USA

[3] Department of Genetics, Developmental and Cell Biology, Iowa State University, Ames, IA 50011, USA

[4] Corn Insects and Crop Genetics Research Unit, USDA-Agricultural Research Service, Ames, IA 50011, USA

[5] Crop Improvement and Genetics Research Unit, USDA-Agricultural Research Service, Albany, CA 94710, USA

[6] Department of Plant Pathology & Microbiology, Iowa State University, Ames, IA 50011, USA

[7] Department of Computer Science, Iowa State University, Ames, IA 50011, USA

*To whom correspondence should be addressed:
Email: Carson.Andorf@usda.gov

Abstract

Background

Gene annotation in eukaryotes is a non-trivial task that requires meticulous analysis of accumulated transcript data. Challenges include transcriptionally active regions of the genome that contain overlapping genes, genes that produce numerous transcripts, transposable elements and numerous diverse sequence repeats. Currently available gene annotation software applications depend on pre-constructed full-length gene sequence assemblies which are not guaranteed to be error-free. The origins of these sequences are often uncertain, making it difficult to identify and rectify errors in them. This hinders the creation of an accurate and holistic representation of the transcriptomic landscape across multiple tissue types and experimental conditions. Therefore, to gauge the extent of diversity in gene structures, a comprehensive analysis of genome-wide expression data is imperative.

Results

We present FINDER, a fully automated computational tool that optimizes the entire process of annotating genes and transcript structures. Unlike current state-of-the-art pipelines, FINDER automates the RNA-Seq pre-processing step by working directly with raw sequence reads and optimizes gene prediction from BRAKER2 by supplementing these reads with associated proteins. The FINDER pipeline (1) reports transcripts and recognizes genes that are expressed under specific conditions, (2) generates all possible alternatively spliced transcripts from expressed RNA-Seq data, (3) analyzes read coverage patterns to modify existing transcript models and create new ones, and (4) scores genes as high- or low-confidence based on the available evidence across multiple datasets. We demonstrate the ability of FINDER to automatically annotate a diverse pool of genomes from eight species.

Conclusions

FINDER takes a completely automated approach to annotate genes directly from raw expression data. It is capable of processing eukaryotic genomes of all sizes and requires no manual supervision – ideal for bench researchers with limited experience in handling computational tools.

Keywords

Genomics, transcriptomics, eukaryotic gene annotation, gene prediction, optimized RNA-Seq alignment, changepoint detection

Background

Recent advances in sequencing technology enable the construction of chromosomal-level assemblies for even non-model organisms. As of December 2020, genomes of 16,108 eukaryotes, 295,784 prokaryotes, 41,936 viruses, 26,079 plasmids and 17,820 organelles are sequenced and available through GenBank [1], a considerable increase over the 1,500 sequences reported two decades ago (see Additional File 1: Figure S1). Therefore, to annotate the ever-rising number of genome sequences, annotation software applications need to be fast, accurate, and designed to handle large amounts of expression data to facilitate discovery of novel genes across different conditions [2–5]. Extensive analysis of this available data is the key to achieving exhaustive gene discovery by analyzing samples from multiple tissues and conditions, obviating the need for additional sequencing.

Genome annotation is the process of identifying transcriptionally active regions of the genome and defining gene structures. Decoding the correct structures of genes is essential since several downstream applications rely on accurate annotations: detecting interactions between proteins [6–14], identifying post-translational modifications [15–23], mining effectors [24–28], and determining protein structure [29–32]. Although we have seen a significant improvement in genome sequencing technology, annotation methods continue to underperform [33, 34]. Obtaining accurate gene annotations is challenging, especially in recently sequenced non-model organisms. The presence of sequences exchanged through horizontal gene transfer in such genomes and the existence of fragmented assemblies make it difficult to predict gene structures [35]. Multiple groups working on the same species have different and oftentimes conflicting annotations that are difficult to merge into a common consensus.

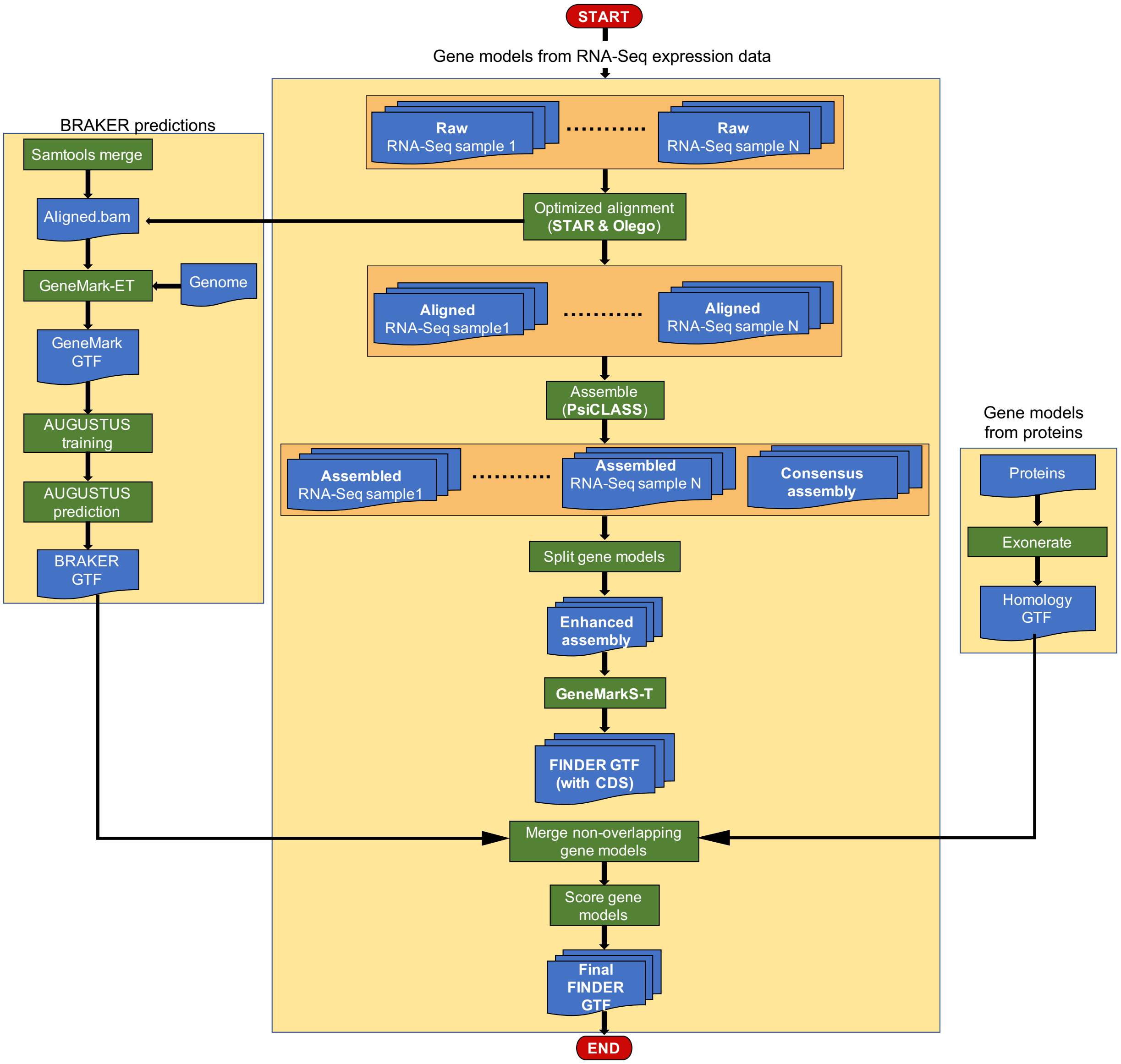
The early 2000s saw initial genome annotation attempts with the introduction of PASA [36], which was developed to map full-length transcripts and Expressed Sequence Tags (ESTs) in order to annotate genomes. In parallel, FGENESH [37, 38], GeneGenerator [39], mGene [40] and GeneSeqer [41] were introduced which predicted gene structures directly from genome sequence. Tools such as MAKER [42–45] and PASA [36] closely depend on pre-assembled full-length transcripts to generate annotations. ESTs and/or *de novo* assembled transcriptomes have been often provided as inputs to these tools to generate annotations [46–52]. Transcripts constructed via *de novo* [53–57] or genome-guided [58–63] approaches are sensitive to the nature of the assembler and its parameter settings. Such assemblers report sequences that are highly similar to one another, making the process of sifting the correct assemblies from artefacts difficult. This issue is moderately mitigated by BRAKER2 [64, 65], which uses read splice information instead of full-length assemblies to predict gene structures and has been shown to perform better than *de novo* approaches [66]. BRAKER2 entails a round of unsupervised gene predictions using GeneMark-ET [67] generating *ab-initio* gene predictions followed by a second round of training by AUGUSTUS [68] using a subset of the gene models created by GeneMark-ET [64]. All variations of MAKER (MAKER, MAKER2 and MAKER-P) use a combination of AUGUSTUS [68] and SNAP [69] to generate gene predictions. Unlike BRAKER2 or PASA, users need to run MAKER for multiple rounds to improve annotation. With no standard technique to optimize the number of rounds, users often undertake a trial-and-error approach to decide what data is supplied to MAKER in each execution round. These unguided choices can create different annotations based on the same data sets. Thus, current approaches report either incomplete genes and/or derive annotations that are missing alternatively spliced transcripts.

To overcome the drawbacks described above, we developed FINDER, a new, automated annotation pipeline that downloads RNA-Seq data from NCBI SRA [70], conducts genome-guided assembly of short reads, predicts gene structure, and annotates genes. FINDER annotates both untranslated and coding regions of genes, categorizes transcripts based on the tissue/conditions where they are expressed, and outputs a complete set of alternatively spliced transcripts. FINDER analyzes the spatial expression profile of each transcript to redefine its boundaries and/or even create newer transcripts and employs an optimized strategy to locate transcripts housing micro-exons. Finally, gene models predicted by BRAKER2 are incorporated into the annotation along with assemblies generated by PsiCLASS [63]. We show that FINDER outperforms state-of-the-art annotation tools in constructing accurate gene structures, when executed with the same expression data.

Implementation

The detailed workflow of FINDER is outlined in **Fig. 1**. The pipeline accepts metadata via a comma-separated values (csv) file (**see Additional file 2: Table S1**). Users can verify the input data using the `verifyInputsToFINDER` utility (Please check section 1.5.1 of Additional file 9). Both single-end and paired-end data are accepted. The pipeline automatically downloads RNA-Seq data from NCBI SRA or the samples can be accessed locally. Multiple rounds of alignment are conducted using STAR [71, 72] with short reads, thus ensuring the capture of tissue-specific splice junctions and ultimately generates the most comprehensive set of alternatively spliced transcripts. FINDER uses PsiCLASS [63] to generate transcripts both at the tissue level and consolidates them to produce a consensus annotation. It employs change-point detection (CPD) using coverage data to polish intron/exon boundaries if needed. Polished transcripts are then supplied to GeneMarkS-T [73] to predict protein coding regions. In addition to constructing genes from expression data, FINDER uses BRAKER2 [65] to predict genes *de novo*. Finally, gene models are assigned scores that reflect the confidence of prediction and evidence across different data sets. Throughout the pipeline run, intermediate temporary data is removed to optimize space usage. Proper logging of executions is implemented through ruffus [74].

Fig 1.



FINDER workflow. FINDER assembles short reads from RNA-Seq expression data, collected from multiple tissues and conditions, to generate full-length transcripts using PsiCLASS. Short read coverage profile is used to polish the structure of the transcripts to enhance the quality of annotation. GeneMarkS-T is used to predict coding regions of the transcripts. Gene models predicted by BRAKER2 and models obtained by mapping proteins are added to the gene models constructed from RNA-Seq data. Additionally, FINDER outputs the tissues where each transcript is expressed allowing users to work with tissue-specific transcripts. FINDER categorizes transcripts into two confidence levels depending on the available supporting evidence and depth of coverage.

Read alignments to the genome

Reads from each sample are aligned to the genome using STAR [72]. FINDER accepts the location of the genomic STAR indices. If indices are not provided, then FINDER will generate them locally. FINDER implements multiple strategies to detect as many correct splice-junctions as possible. Several studies use a multi-step approach where splice junctions are detected in the first pass and then those junctions are used to guide the alignments in future passes [75, 76]. FINDER employs a similar strategy to align reads and obtain the most confident splice junctions in each tissue type and/or condition by conducting mapping in four passes (Please check section 1.3 of Additional file 9 for more details).

Annotating transcripts with micro-exons

Certain genes in eukaryotes have micro-exons (i.e., exons with fewer than 50 nucleotides) [77–80] which impart important biological properties both in plants [81–85] and animals [86–90]. FINDER uses OLego [91] to map the reads which were reported unmapped by STAR, because OLego optimizes micro-exon sensitivity by checking intron signatures when no hits of seed sequences (~14 nt) are found. It is configured to align reads to exons of minimum length 2, with a minimum and maximum intron size of 20 and 10K respectively.

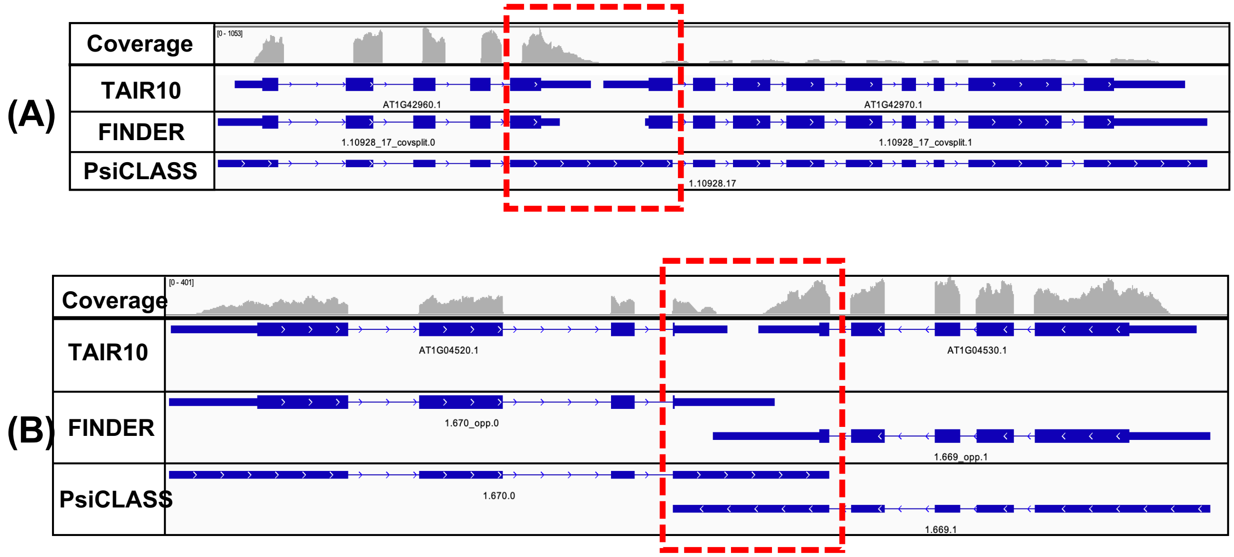
Generating exon-exon transcript structure annotation with PsiCLASS

Alignments reported by STAR and OLego are combined and provided as input to PsiCLASS [63]. Unlike traditional assemblers, PsiCLASS accepts alignments from multiple samples at the same time. It generates annotations for each sample and one consolidated gene annotation for all the samples. FINDER runs PsiCLASS with the `--bamGroup` option enabled which instructs PsiCLASS to preserve tissue/condition specific features. It is a fast meta-assembler generating 350 samples of output in less than three hours while running on 30 cores and consumes less than 50 GB of memory.

Polishing gene structures to optimize gene discovery

Gene structure annotations reported by PsiCLASS were polished to generate the best assemblies. Annotations generated by assemblers often have three kinds of errors that impact accuracy: (1) presence of redundant transcripts that are proper subsets of other transcripts, (2) multiple transcripts on the same strand merged into one, and (3) transcripts with ill-defined exon boundaries. Most assemblers ignore such cases to boost the speed of operation. Developing solutions to deal with these kinds of errors increases the number of correct structural annotations thereby improving downstream analysis.

Fig 2.



FINDER implements changepoint analysis of read coverages to modify existing gene models and/or generate new ones. Changepoint analysis is a statistical technique to assess alterations in trends over time. The same approach has been used to analyze read coverage patterns of a genome, where the data is distributed spatially. (A) Two *Arabidopsis thaliana* genes AT1G42960.1 and AT1G42970.1 are present within 50 base pairs of each other on the positive strand. Reads originating from the end exons of either genes bleed into each other resulting in PsiCLASS to merge the two gene models. Changepoint analysis recognizes the fall the read coverage and reports a position within the exon where the trough exists. This information is used to split up the gene models. (B) A similar issue exists with closely spaced genes residing on opposite strands. The end exons (highlighted with a red box) for a transcript extend up to the nearest intron of the adjacent transcript. Changepoint analysis is used to determine the actual end/start of transcript based on the read coverage.

FINDER uses different algorithmic and statistical approaches to deal with the above cases. To eliminate redundant transcripts, exon-intron structure of all transcripts is compared with each other to retain only unique transcripts. Even though eukaryotes possess large genomes, certain genes/transcripts are closely packed and are overlapping (Fig. 2). Reads originating from one of those genes often map to nearby overlapping genes making the task of distinctly recognizing the transcripts very challenging.

FINDER is configured to use changepoint detection (CPD) analysis to detect the descent in read coverage at the junction of two overlapping transcripts. Statistical CPD is a procedure to detect changes in the probability distribution of a stochastic process. Typically, CPD is widely used to detect changes in time series [92–96], but can be extended to other applications as well [97, 98]. We have found that even though CPD was developed under the assumption of normality, it can also be used where normality is violated.

In the first step in FINDER's CPD, short read alignments to the genome are converted into number of read counts per nucleotide using bedtools [99]. A custom python script is used to transfer the per nucleotide coverage data from the genome to the transcriptome reported by PsiCLASS. Each internal exon is considered as a potential site for the presence of changepoints if there exist premature stop codons in all the three frame translations. CPD only considers exons that have a high chance of housing a changepoint, thereby reducing duration of operation. The coverage pattern of each exon is probed to detect changepoints. The data has been modeled using an exponential distribution, and binary segmentation has been used to determines the changepoints in the exonic coverage using the 'changepoints' package [100]. Read coverage of exons mimics a time series where each nucleotide position of an exon can be assumed to be a single unit of time. Coverage patterns of exons, suspected to be merged, contain a characteristic depression in the signal to split the gene models (**Fig. 2A**). Overlapping transcripts on opposite strands sometimes share a common exon (**Fig. 2B**). This negatively impacts precision since the boundaries of the predicted transcript exceed the boundaries of the transcript in the reference annotation. FINDER trims the transcript boundaries, using the changepoints, to better model the RNA-Seq coverage (**Fig. 2B**). These strategies improve the annotation by increasing the transcript F1 scores (**Table 1**).

Table 1: Sensitivity, Specificity and F1 scores of transcripts generated by multiple gene annotation pipelines for three model organisms – *Arabidopsis thaliana, Oryza sativa* and *Zea mays*

	Arabidopsis thaliana							*Oryza sativa*							*Zea mays (NCBI)*					
	BRAKER2	MAKER2	PASA	FINDER	FINDER+BRAKER2	FINDER+BRAKER2+PROTEIN		BRAKER2	MAKER2	PASA	FINDER	FINDER+BRAKER2	FINDER+BRAKER2+PROTEIN		BRAKER2	MAKER2	PASA	FINDER	FINDER+BRAKER2	FINDER+BRAKER2+PROTEIN
Base Specificity	**91.08**	74.87	62.71	74.46	75.01	75.04		**57.63**	52.67	36.77	42.46	42.6	42.75		6.52	45.22	60.39	**62.97**	**62.96**	**62.92**
Base Sensitivity	60.27	52.55	69.41	**71.45**	**74.14**	**74.23**		36.85	40.17	59.19	**61.43**	**61.83**	**62.37**		50.43	54.38	66.7	**72.1**	**72.15**	**72.16**
Base F1 score	72.54	61.76	65.89	**72.92**	**74.57**	**74.63**		44.95	45.58	45.36	**50.21**	**50.44**	**50.73**		11.55	49.38	63.39	**67.23**	**67.24**	**67.22**
Exon Specificity	80.28	**95.35**	90.74	91.79	91.43	91.43		38.4	**72.31**	67.37	67.67	67.5	67.62		15.96	65.14	**84.26**	79.71	79.59	79.57
Exon Sensitivity	73.29	55.04	67.49	**69.93**	**71.62**	**71.7**		53.74	50.59	63.99	**64.81**	**65.22**	**66.06**		65.65	61.84	71.84	**74.38**	**74.4**	**74.4**
Exon F1 score	76.63	69.79	77.41	**79.38**	**80.32**	**80.37**		44.79	59.53	65.64	**66.21**	**66.34**	**66.83**		25.68	63.45	**77.56**	76.95	76.91	76.9
Intron Specificity	86	**98.26**	95.98	96.55	96.26	96.26		52.8	**76.87**	72.36	73.24	73.19	73.26		21.77	74.41	**88.45**	84.86	84.78	84.76
Intron Sensitivity	84.25	60.64	76.11	**76.51**	**78.33**	**78.42**		**76.78**	55.77	70.44	71.03	71.52	72.34		80.88	64.99	75.19	**78.12**	**78.15**	**78.15**
Intron F1 score	85.12	75	84.9	**85.37**	**86.37**	**86.43**		62.57	64.64	71.39	**72.12**	**72.35**	**72.8**		34.31	69.38	**81.28**	**81.35**	**81.33**	**81.32**
Transcript Specificity	49.91	**76.9**	55.21	60.04	59.82	59.82		12.71	**40.13**	23.17	24.54	24.59	24.54		2.75	32.67	**48.6**	44.78	44.84	44.78
Transcript Sensitivity	30.26	21.74	28.62	**37.21**	**39.21**	**39.28**		16.35	18.01	29.51	**33**	**33.32**	**33.82**		19.57	25.01	37.94	**42.25**	**42.3**	**42.31**
Transcript F1 score	37.68	33.9	37.7	**45.95**	**47.37**	**47.42**		14.3	24.86	25.96	**28.15**	**28.3**	**28.44**		4.82	28.33	42.61	**43.48**	**43.53**	**43.51**
Gene Specificity	51.58	**76.9**	61.79	67.33	66.56	66.55		13.46	**40.13**	36.81	32.79	32.76	32.38		2.78	32.67	**55.69**	52.3	52.29	52.19
Gene Sensitivity	50.43	37.2	44.39	57.96	**61.3**	**61.4**		18.07	19.4	31.37	**35.05**	**35.4**	**35.96**		29.02	38.14	51.57	**57.35**	**57.42**	**57.43**
Gene F1 score	51	50.14	51.66	**62.29**	**63.82**	**63.87**		15.43	26.16	33.87	**33.88**	**34.03**	**34.08**		5.07	35.19	53.55	**54.71**	**54.74**	**54.68**

De novo gene prediction from expression data and proteins from closely related species

Certain genes are expressed only under specific tissues and conditions [101]. However, constructing an exhaustive set of genes expressed across all possible tissues and conditions is a daunting task due to the mammoth volume of potential expression data. Hence, approaches that can predict structures of unknown genes using information obtained from known genes are needed. Within the FINDER framework, we used BRAKER2 [65] to predict the structure of protein coding genes. The pipeline is provided with alignment files generated by STAR and an optional, user-provided protein data file. If the previous execution fails, a second execution of BRAKER2 is launched without protein information. Genes predicted by BRAKER2 are compared to the genes obtained from expression data. To prevent too many false positives, predictions made by BRAKER2 are considered high confidence, only if those are supported by expression level or protein level evidence.

Prediction of coding regions

We leveraged GeneMarkS-T [73] to predict protein-coding regions of genes constructed from expression data. GTF files are first converted to FASTA files using the provided genome. Those FASTA files are supplied to GeneMarkS-T as inputs. GeneMarkS-T outputs coding sequence for the transcripts. CDS annotations are incorporated into the final GTF file by converting the transcriptomic coordinates to genomic coordinates.

Using proteins to annotate more genes

In addition to RNA-Seq data, FINDER also uses protein data (when provided), in two ways (1) to assess the veracity of the transcript models generated by BRAKER2, and (2) to align those proteins not recognized by BRAKER2 or PsiCLASS. Protein coding genes obtained from expression data and predicted by BRAKER2 are BLASTed [102] to the protein set provided by the user. Proteins not encountering any hits are aligned to the genome using exonerate [103] with a minimum threshold of 90% similarity. These alignments are augmented to the final set of gene predictions. Since these transcripts are obtained solely from proteins, they lack UTR sequences.

Tissue/condition specific transcripts/gene models

Most eukaryotic genes have multiple isoforms which are derived from alternative transcripts. Expression of different transcripts can occur under different conditions in different tissues at different time points. FINDER compares assembled transcripts from each condition and prints out an association between each transcript and the provided tissue/condition (Additional file 9 section 1.5).

Scoring gene models

FINDER groups genes into multiple categories based on supporting evidence. Genes that are expressed in RNA-Seq datasets, predicted by BRAKER2, and have protein evidence, are put into the high-confidence gene set. BRAKER2-predicted genes with no evidence of expression and/or proteins are treated as low confidence genes. FINDER expects a soft masked genome since it is a BRAKER2 requirement. Genes which are located in the repeat regions are marked as such and moved to the set of low-confidence genes.

Results & Discussion

Choice of species for comparison

We tested the performance of FINDER primarily on three well-annotated plant organisms - *Arabidopsis thaliana* [104]*, Oryza sativa* [105–107] and *Zea mays* [108, 109]*.* The genomes assemblies of these model organisms have been frequently updated and are almost complete with telomere-to-telomere sequences with fewer gaps and unknown nucleotides. In addition, their gene annotations have undergone regular improvement by mining the large number of RNA-Seq datasets available in the literature. Also, The Arabidopsis Information Resource (TAIR) provides a five-star rating system based on available evidence for each gene. Such a system offers a platform to test the quality of gene annotation software. For further evaluation, and to ensure that FINDER is able to annotate a wider range of genome types, we selected the following additional species to test: C*aenorhabditis elegans* [110]*, Drosophila melanogaster* [111, 112]*, Homo sapiens* [113, 114]*, Hordeum vulgare* [115]*,* and *Triticum aestivum* [115–118]). The genomes of these species range from small (*C. elegans*, *D. melanogaster, A. thaliana*), medium (*O. sativa*), to large (*H. sapiens, Z. mays*, *H. vulgare,* and *T. aestivum*). Finally, we evaluated FINDER on three different versions of *Z. mays* annotations – RefSeq [119], AGPv3 [109, 120] and AGPv4 [108, 121].

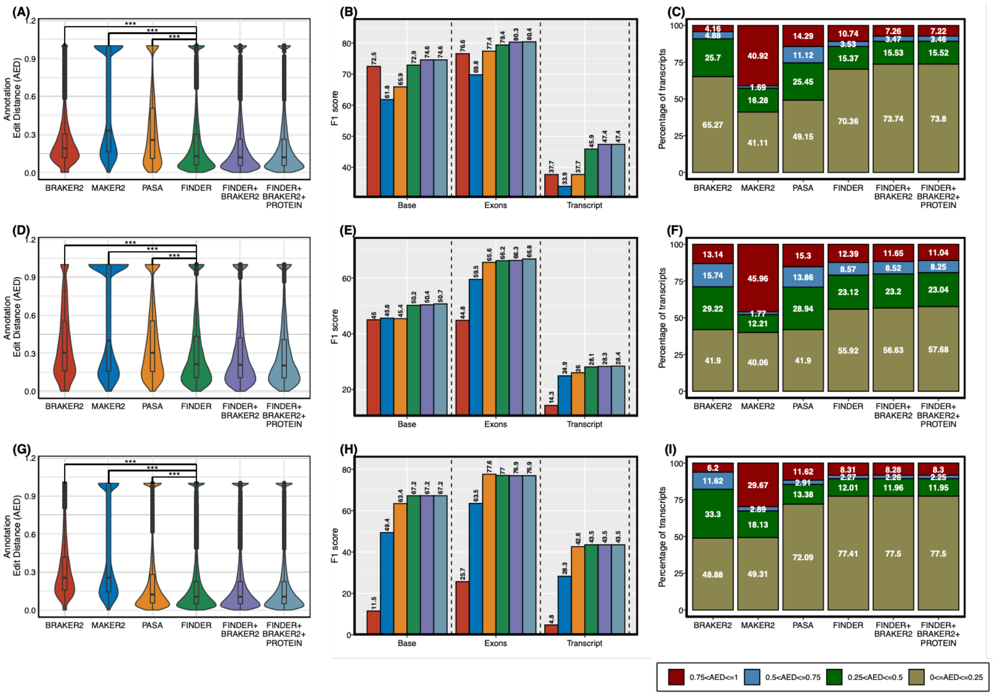
Metrics to assess quality of annotation

We used four metrics to compare the quality of annotations generated by each pipeline: 1) Annotation Edit Distance (AED) [42, 43, 122], 2) sensitivity, 3) specificity, and 4) F1 score. Although these metrics could be computed both at the nucleotide- and exon-level we chose to make comparisons at the transcript level since it encompasses bases, exons, and introns. An AED score of 0 indicates complete agreement of the predicted annotation with the reference, and a score of 1 denotes that the reference has not been identified in the annotation. A transcript is considered to be “recognized” only when all its intron definitions agree with at least one transcript from the predicted set. We used the Mikado “compare” utility to compare the predictions with the reference annotations [123]. A highly sensitive annotation is one that can correctly recognize more reference transcripts. A set of annotations has high specificity when it reports minimal incorrect transcripts. For an annotation to be of good quality, both sensitivity and specificity should be high. A balanced metric is the F1 score which is the harmonic mean of sensitivity and specificity. While AED provides a good numeric assessment of how well the ground truth evidence is represented in an annotation, when individually used, it fails to capture the extent to which false positives are reported. Hence, F1 score complements AED since it incorporates both specificity and sensitivity. For evaluation purposes, we assume that the annotations achieved through community efforts are the ground truth and contain no errors.

FINDER generates more accurate gene models than BRAKER2, MAKER2 and PASA

FINDER leverages expression data to construct transcript models and employs statistical changepoint detection to enhance their structures (see Implementation). Both MAKER2 and PASA were run with transcript sequences reported by PsiCLASS.

Fig 3.



Comparison of performance of predicted annotations in three model species – (A – C) *A. thaliana*, (D – F) *O. sativa* and (G – I) *Z. mays.* Annotation Edit Distance (AED) is an assessment of how well predicted annotations agree with the evidence and was used as a quality control metric. A value of 0 denotes complete agreement of two annotations while a value of 1 denotes that the 'gold standard' reference annotation was not detected. Transcripts from 'gold standard' reference annotations that are not detected in any of the predicted annotations are removed from analysis. (**A, D & G**) Distribution of AED scores. Violin plots wider at the base indicate high density of annotations with lower AED. FINDER was able to create gene models having lowest AED resulting in a wide base. Gene models generated by FINDER were enhanced by adding predictions made by BRAKER and including protein evidence. Wilcoxon's signed rank test was used to compare the AED scores between FINDER and other annotating pipelines. The "***" symbol implies that the AED scores of FINDER gene models were significantly lesser (p_value<0.01) than the AED scores of the gene models reported by other pipelines. (**B, E & H**) Bar plot of F1 score of multiple approaches of annotation. Having a high nucleotide F1 (Base F1) or a high exon F1 score is not sufficient to conclude a good annotation. High value of transcript F1 score is indicative of good gene models with high sensitivity and high specificity. (**C, F & I**) Stacked bar plot showing percentage of transcripts in each of the four groups of AEDs. Higher number of transcripts to low AED denotes better annotation. In each of the three species, FINDER was able to generate a higher percentage of transcripts with low AED compared to other techniques of annotation.

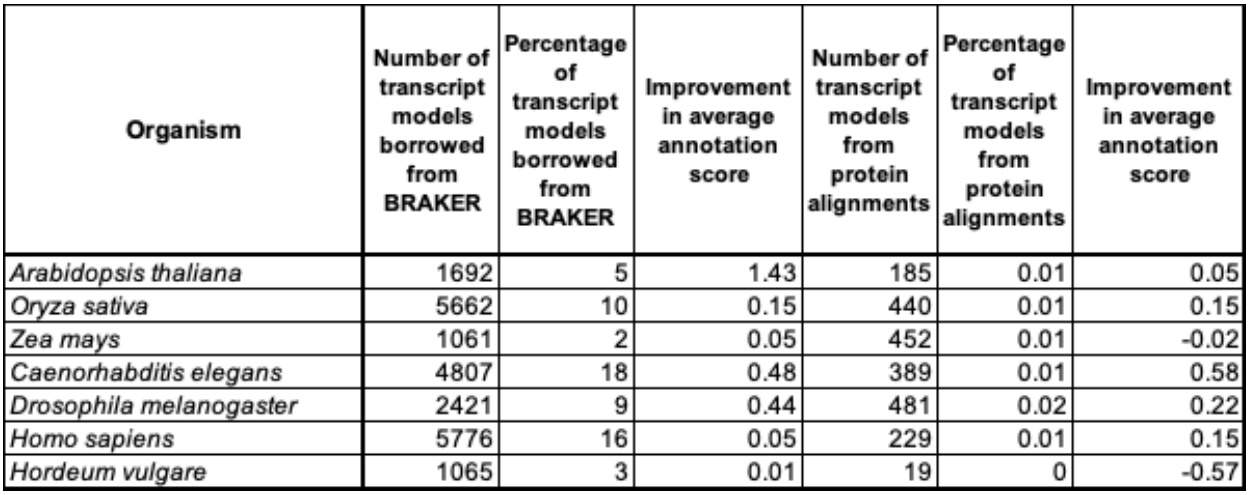
To assess FINDER's performance, we compared the AED scores of transcript models generated by FINDER with those generated by other commonly used annotation methods. As shown in **Fig. 3A, 3D & 3G,** the violin plots for FINDER are broader at the base, indicating a greater number of transcripts with lower AED scores as compared to BRAKER2, MAKER, and PASA. We compared the FINDER AED scores with the AED scores reported by other pipelines using Wilcoxon's signed rank test (More details in Additional file 9 section 2.5). For all organisms (Fig. 3, **Additional file 1: Figure S2–S5 and Additional file 3: Table S2**), the AED scores reported by FINDER were significantly lower (p_value<0.01) than that of any other pipeline. **Fig. 3C, 3F & 3I**, shows a stacked bar plot to represent the fraction of transcripts in each category of AED values. In all the cases, a higher percentage of transcripts reported by FINDER have lower AED scores (**Additional file 1: Figure S2-S5**). This indicates that FINDER is capable of constructing gene structures that better comply with the reference annotations.

High-quality exhaustive annotations predict the fewest false positives thereby boosting the transcript F1 score. The transcript F1 scores of the gene models that were reported by FINDER for *A. thaliana*, *O. sativa* and *Z. mays* were higher than the models generated by BRAKER2, MAKER, and PASA (**Fig. 3B, 3E & 3H**). This same trend is observed for other tested organisms where FINDER was successful in detecting nucleotides, exons, introns, transcripts and genes (**Table 1, Additional file 1: Figure S2-S5 and Additional file 3: Table S2**). MAKER2 and BRAKER2 registered a high specificity for most of the organisms because fewer transcripts were reported than FINDER. MAKER2 and BRAKER2 also had lower F1 scores, indicating less sensitivity than FINDER. Additionally, we compared the CDS regions of genes reported by FINDER with those of BRAKER2. For most of the organisms, FINDER generated transcript models with a higher F1 score (**Additional file 4: Table S3**). These results show that the better performance of FINDER is ensured not only due to the presence of UTRs but also due to enhanced CDS structure of gene models.

Finally, including BRAKER2 predictions and protein sequences to FINDER enhanced the gene model predictions. About 15% of the gene models reported by BRAKER2, those having high sequence similarity with the provided protein sequences were included in the final annotations (**Table 2**). As shown in **Table 1 and Additional file 5: Table S4**, including evidence at the protein level led to the identification of more genes.

Unlike BRAKER2, FINDER does not assume a homogeneous nucleotide composition of the genome [124]. FINDER outperforms BRAKER2 while constructing gene models in complex organisms like *H. sapiens, H. vulgare,* and *Z. mays* since assemblers generating transcriptomes from alignments do not require a genome to possess homogeneous nucleotide composition.

Table 2: Improvement in overall gene recognition by adding gene models predicted by BRAKER2 and aligning protein sequences



Organism	Number of transcript models borrowed from BRAKER	Percentage of transcript models borrowed from BRAKER	Improvement in average annotation score	Number of transcript models from protein alignments	Percentage of transcript models from protein alignments	Improvement in average annotation score
Arabidopsis thaliana	1692	5	1.43	185	0.01	0.05
Oryza sativa	5662	10	0.15	440	0.01	0.15
Zea mays	1061	2	0.05	452	0.01	-0.02
Caenorhabditis elegans	4807	18	0.48	389	0.01	0.58
Drosophila melanogaster	2421	9	0.44	481	0.02	0.22
Homo sapiens	5776	16	0.05	229	0.01	0.15
Hordeum vulgare	1065	3	0.01	19	0	-0.57

FINDER in itself is restricted to annotate genes only in regions of the genome that are transcriptionally active. Recognizing that BRAKER2, being a gene predictor, can construct gene models in transcriptionally silent regions of the genome, FINDER is designed to incorporate the gene models predicted by BRAKER2 into the final annotations.

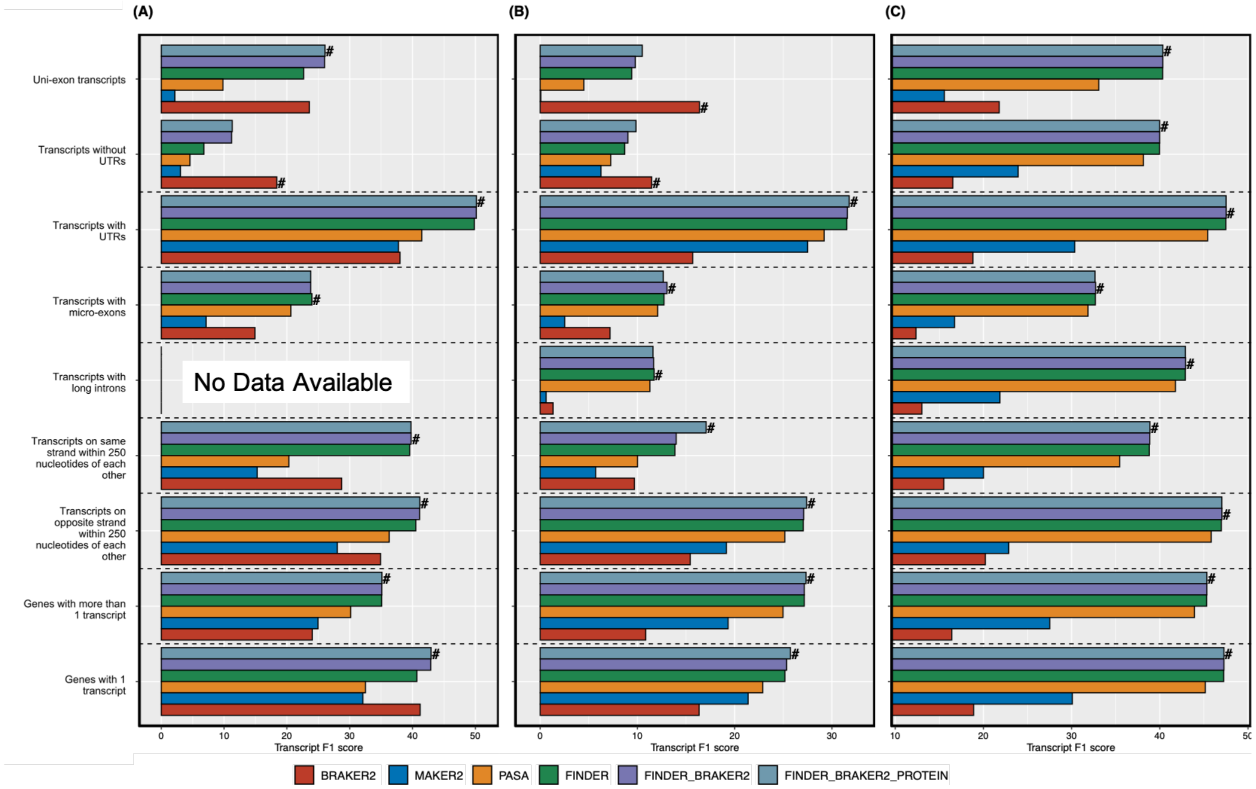
Table 3 Classification of gene models into different groups based on their relative location to other genes, number of isoforms and other criteria.

	Name	Description
Group 1	Uni-exon transcripts	Transcripts having a single exon and no introns
Group 2	Transcripts without UTRs	Transcripts missing either the 5' or the 3' UTR sequence
Group 3	Transcripts with UTRs	Transcripts having both UTRs
Group 4	Transcripts with micro-exons	Transcripts where at least one exon has length less than 50 nucleotides
Group 5	Transcripts with long introns	Transcripts where at least one intron has a length greater than 10,000 bp
Group 6	Closely placed transcripts on same strand	Transcripts on the same strand having less than 250 nucleotides between each other
Group 7	Closely placed transcripts on opposite strand	Transcripts on the opposite strands having less than 250 nucleotides between each other
Group 8	Multi transcript gene	Transcripts of a gene that have multiple transcripts
Group 9	Single transcript gene	Transcripts of a gene that have single transcript

Distinct gene groups are accurately annotated with FINDER

Although eukaryotic genes differ from one another in terms of location, structure and the isoforms they encode, most annotation pipelines annotate and evaluate gene predictions with a global and uniform approach. The problem arises when these variances prompt each pipeline to perform differently on dissimilar groups of genes. To avoid this pitfall, we created groups of genes and transcripts based on various criteria (**Table 3)** and compared the performance of FINDER with BRAKER2, MAKER, and PASA for each of these sets.

Fig 4.



FINDER vs other pipelines on different groups of genes in three model species – (A) *A. thaliana*, (B) *O. sativa* and (C) *Z. mays*. F1 score is the harmonic mean between sensitivity and specificity. Higher F1 score indicates better agreement with the reference transcript models. We created groups of transcripts that have similar characteristics as shown in the y-axis legend. A pool of transcripts was created containing multi-exonic transcript predictions, from each pipeline, that has a complete intron chain match with at least one reference annotation. Mono exonic transcripts were considered if at least 80% of the nucleotides overlap with one reference annotation. Transcript F1 scores, for each of the annotation pipelines, have been plotted as a bar graph. Even though all annotation pipelines are designed to serve the same purpose of annotating genomes, each pipeline adopts a different strategy. Each strategy has its own merits and demerits that lead to better annotation of a certain category of genes. This plot helps understand the performance of each annotation pipeline on different categories. The symbol "#" denotes the best annotator in each gene group.

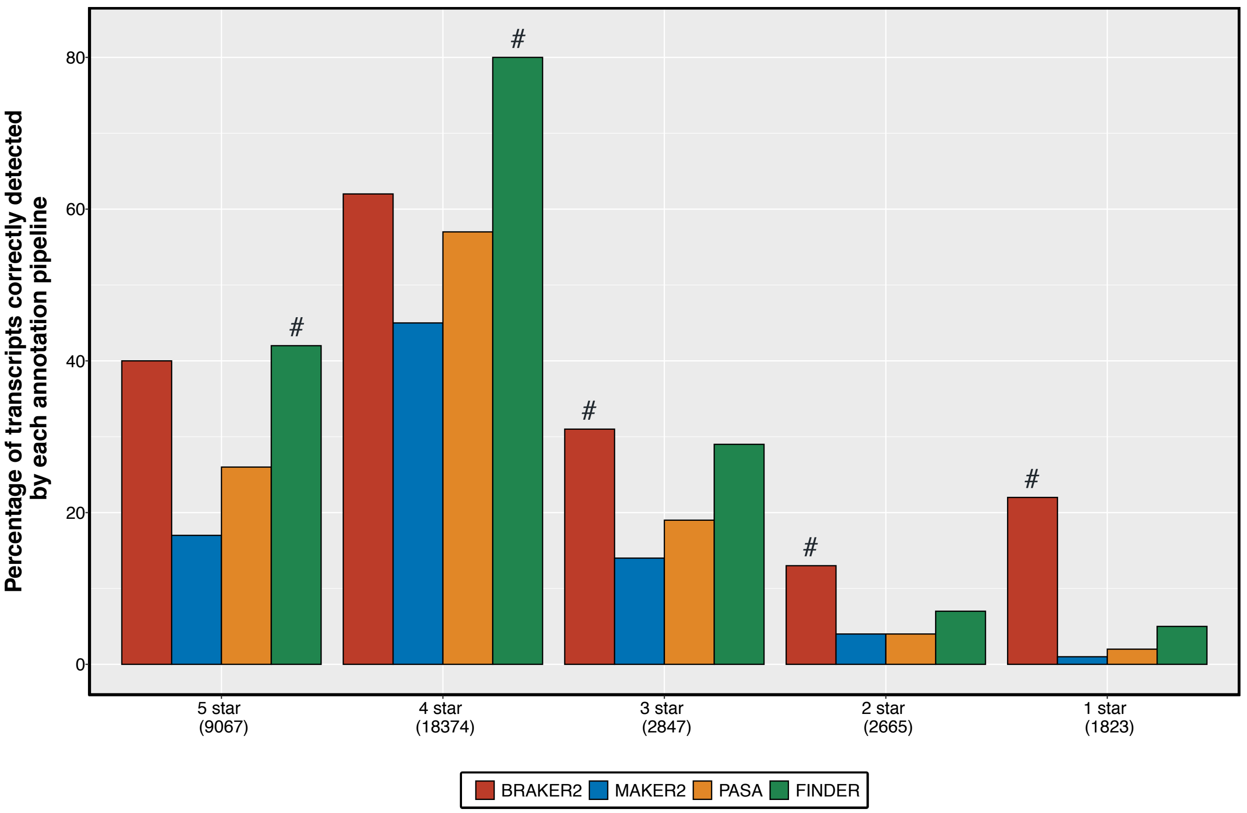
On the set of UTR-containing transcripts, FINDER reported the best transcript F1 scores (**Fig. 4, Additional file 1: Figure S6, S7**). Unlike BRAKER2, FINDER uses GeneMark S/T to predict CDS from the transcript sequences assembled by PsiCLASS and can hence annotate UTR regions. For most of the organisms, BRAKER2 and MAKER2 gene models register a low transcript F1 score in this category of genes. Next, we tested the performance of the annotation pipelines on transcripts that are closely located in the genome. On this set of transcripts, FINDER reported the best F1 transcript score for *A. thaliana, O. sativa,* and *Z. mays* (**Fig. 4**), and comparable scores for *D. melanogaster* (**Additional file 1: Figure S6**), *H. vulgare* (**Additional file 1: Figure S8**), and *C. elegans* (**Additional file 1: Figure S7**) with BRAKER2. Most eukaryotic genes have multiple isoforms which differ from one another by their exon-intron definition. Splice sites and coverage information provides clues to construct such alternatively spliced transcripts. We selected genes with more than one transcript to check how well each annotation pipeline was able to detect transcript isoforms. For this case, FINDER was able to generate the best transcript structures with the highest transcript F1 score among all the pipelines gene annotation software applications (**Fig. 4 and Additional file 1: Figure S6-S9).** Surprisingly, BRAKER2 fared poorly in this category despite training with all the detected splice sites from RNA-Seq data. This demonstrates that FINDER is capable of leveraging both intron splice sites and read coverages to report best transcript structures. For *H. sapiens*, PASA was able to generate the best transcript structures across all categories of transcripts. Adding transcripts from BRAKER2 and protein evidence improved the transcript F1 score for all the organisms, signifying the importance of incorporating *de novo* gene models and protein evidence.

BRAKER2 generated the best transcript annotation for the set of transcripts with a single exon (**Fig. 4A&B** and **Additional file 1: Figure S6-S9**). Such transcripts, devoid of any introns, are difficult to construct from RNA-Seq alone. Also, the direction of the splice sites infers the direction of a transcript. Without any introns, such a single-exon transcript has to be probed for a CDS sequences' presence to infer directionality. BRAKER2 was configured to optimally predict only CDS regions of genes, hence, it performs well with the set of transcripts that have missing UTRs for organisms with small and moderate sized genomes (**Fig. 4A&B and Additional file 1: Figure S6-S9).** The average number of transcripts per gene reported by BRAKER2 is lower than FINDER. While this boosts specificity, it compromises recall since BRAKER2 is not sensitive to detecting alternatively spliced transcripts. Hence, BRAKER2 accomplishes the best F1 score when tested on a set of single-transcript genes but performs poorly on a set of multi-transcript genes (**Fig. 4A&B and Additional file 1: Figure S6-S9**).

Performance comparison on TAIR's 5-star System

In order to assess the performance of the annotation pipelines on groups of genes constructed from varying levels of evidence, we used the TAIR10 5-star system. TAIR associates a quality score to each *A. thaliana* transcript based on the evidence used to construct the models, with five stars designating the best evidence and zero stars the least [125]. The three categories with limited evidence (<3 stars) have fewer than 3,000 transcripts each. BRAKER2's performance, on the genes in these three categories, was slightly better than the rest of the annotation pipelines (**Fig. 5**). The other two categories (five star and four star) have 9,067 and 18,374 transcripts respectively. In both of these categories, FINDER was able to detect more transcripts than any other annotation pipeline. 51.5% and 86.4% of genes in the 5-star and 4-star category respectively were multi-exonic. In both these categories, FINDER correctly constructed more gene models compared to any other annotation pipeline (**Fig. 5**). FINDER reported 80% of the gene models belonging to the 4-star category – 18% more than BRAKER2 (**Fig. 5**). Hence, it is evident from this analysis that FINDER can reconstruct the structures of most of the genes that are well-supported by underlying evidence.

Fig 5.

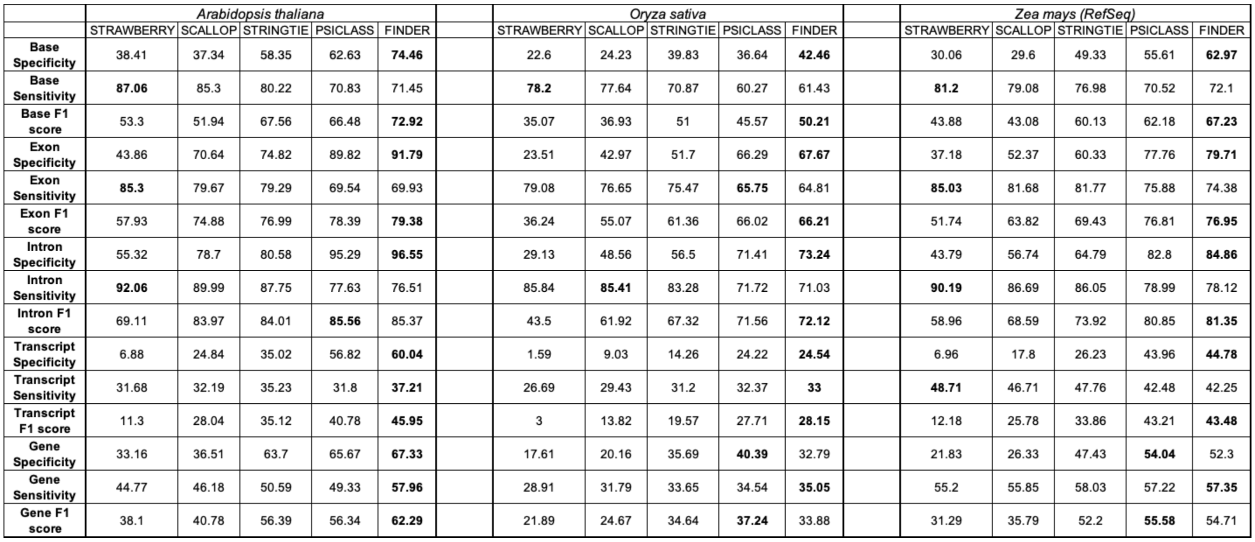


Performance of annotation pipelines on gene groups of *Arabidopsis thaliana* generated by TAIR10. The Arabidopsis Information Resource (TAIR) group has created a quality ranking system to indicate the level of confidence in an annotated gene/transcript. The ranking system has five levels (denoted by stars). Higher number of stars denote the availability of more information to generate the gene structure. Here we display the percentage of transcripts in each category that was identified by a particular annotation pipeline. A high percentage of identified transcripts indicate higher sensitivity and hence a better annotation. The number below each legend in the x-axis denote the number of genes in that respective group. The "#" denotes the predictor which detected the maximum number of transcripts within each group.

Improving transcript annotations using changepoint analysis

The co-location of multiple overlapping genes on the genome strands makes it difficult to correctly annotate their structures (see Methods **Polishing gene structures to optimize gene discovery**). FINDER employs changepoint detection (CPD) [100] to split the merged transcripts reported by PsiCLASS (**Fig. 2**). To gauge the magnitude of improvement in transcript structures brought about by the application of CPD, we compared the accuracy of the predicted transcriptome before and after implementing CPD based on read coverage. As shown in **Table 4 and Additional file 6: Table S5**, implementing the CPD improved both specificity and sensitivity in organisms with small or medium-sized genomes. In *A. thaliana*, the transcript F1 scores increased from 40.78 to 45.95 (**Table 4 and Additional file 6: Table S5**) and in *C. elegans* it increased from 40 to 50. In large genomes, the improvement was not as significant, mainly because there are only a few genes that overlap with one another.

Table 4: Comparison of specificity, sensitivity and F1 scores of transcripts assemblies generated by Strawberry, Scallop, Stringtie, PsiCLASS and FINDER for three model organisms – *A. thaliana, O. sativa* and *Z. mays*.



	Arabidopsis thaliana						*Oryza sativa*						*Zea mays (RefSeq)*				
	STRAWBERRY	SCALLOP	STRINGTIE	PSICLASS	FINDER		STRAWBERRY	SCALLOP	STRINGTIE	PSICLASS	FINDER		STRAWBERRY	SCALLOP	STRINGTIE	PSICLASS	FINDER
Base Specificity	38.41	37.34	58.35	62.63	**74.46**		22.6	24.23	39.83	36.64	**42.46**		30.06	29.6	49.33	55.61	**62.97**
Base Sensitivity	**87.06**	85.3	80.22	70.83	71.45		**78.2**	77.64	70.87	60.27	61.43		**81.2**	79.08	76.98	70.52	72.1
Base F1 score	53.3	51.94	67.56	66.48	**72.92**		35.07	36.93	51	45.57	**50.21**		43.88	43.08	60.13	62.18	**67.23**
Exon Specificity	43.86	70.64	74.82	89.82	**91.79**		23.51	42.97	51.7	66.29	**67.67**		37.18	52.37	60.33	77.76	**79.71**
Exon Sensitivity	**85.3**	79.67	79.29	69.54	69.93		79.08	76.65	75.47	**65.75**	64.81		**85.03**	81.68	81.77	75.88	74.38
Exon F1 score	57.93	74.88	76.99	78.39	**79.38**		36.24	55.07	61.36	66.02	**66.21**		51.74	63.82	69.43	76.81	**76.95**
Intron Specificity	55.32	78.7	80.58	95.29	**96.55**		29.13	48.56	56.5	71.41	**73.24**		43.79	56.74	64.79	82.8	**84.86**
Intron Sensitivity	**92.06**	89.99	87.75	77.63	76.51		85.84	**85.41**	83.28	71.72	71.03		**90.19**	86.69	86.05	78.99	78.12
Intron F1 score	69.11	83.97	84.01	**85.56**	85.37		43.5	61.92	67.32	71.56	**72.12**		58.96	68.59	73.92	80.85	**81.35**
Transcript Specificity	6.88	24.84	35.02	56.82	**60.04**		1.59	9.03	14.26	24.22	**24.54**		6.96	17.8	26.23	43.96	**44.78**
Transcript Sensitivity	31.68	32.19	35.23	31.8	**37.21**		26.69	29.43	31.2	32.37	**33**		**48.71**	46.71	47.76	42.48	42.25
Transcript F1 score	11.3	28.04	35.12	40.78	**45.95**		3	13.82	19.57	27.71	**28.15**		12.18	25.78	33.86	43.21	**43.48**
Gene Specificity	33.16	36.51	63.7	65.67	**67.33**		17.61	20.16	35.69	**40.39**	32.79		21.83	26.33	47.43	**54.04**	52.3
Gene Sensitivity	44.77	46.18	50.59	49.33	**57.96**		28.91	31.79	33.65	34.54	**35.05**		55.2	55.85	58.03	57.22	**57.35**
Gene F1 score	38.1	40.78	56.39	56.34	**62.29**		21.89	24.67	34.64	**37.24**	33.88		31.29	35.79	52.2	**55.58**	54.71

PsiCLASS meta- assembly works better than other approaches

We explored three popularly used software applications for merging transcriptome assemblies – StringTie-merge [76, 126–132], TACO [133–138] and Cuffmerge [139–144] to combine 116 *A. thaliana* assemblies constructed by StringTie [59], Scallop [61] and Strawberry [60] (Please check section 3 of Additional file 9 for more details). The best assembly was reported by StringTie-merge and was hence used for all other organisms. We compared the accuracy of the consensus transcript models generated by StringTie-merge with the transcript models reported by PsiCLASS [63]. As depicted in **Table 4 and Additional file 6: Table S5**, PsiCLASS generated the best transcript models for all organisms registering the highest transcript F1 score improving upon the StringTie models by up to 15%. Hence, FINDER uses only PsiCLASS to generate assemblies from short-read data.

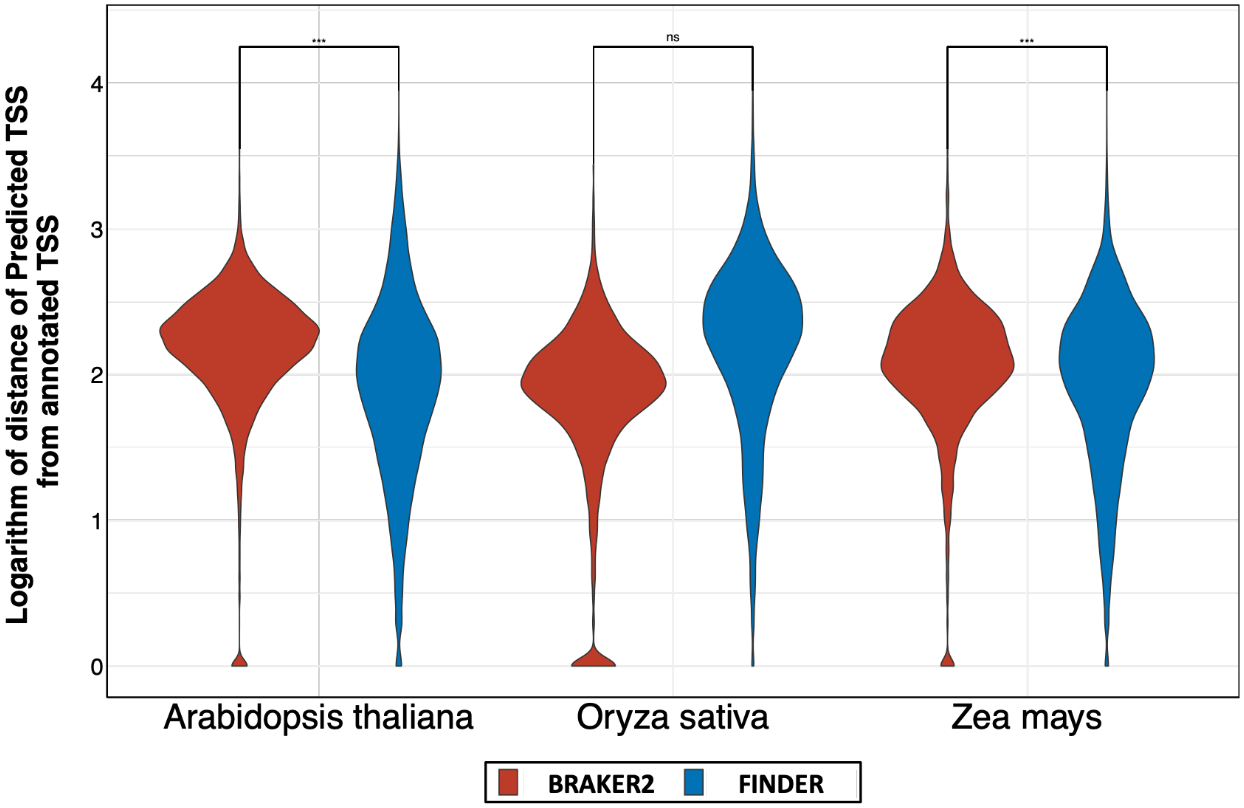
Impact of missing untranslated region on annotation of transcripts

Gene transcription is triggered by adherence of a transcription factor in the promoter region of a gene. Promoters are typically located within 1,000 bp upstream of a gene's transcription start site (TSS) [145–147]. Determining the TSS from sequencing data is best facilitated by RAMPAGE [148, 149] or CAGE-Seq [150], but this data is usually unavailable due to constraints imposed by cost and time. Nevertheless, a good estimate can be obtained from RNA-Seq data by assuming the start coordinates of the assembled genes as the TSS. Thus, researchers often localize their investigation to a section 500-1,000 bp upstream of the assumed TSS [151, 152]. Without 5' UTR annotation it is impossible to deduce a good approximation of the TSS. This leads to conducting promoter mining in a completely incorrect genome location. To assess the quality of 5' UTR annotation, we plotted the difference of TSS between the reference genes and the genes reported by BRAKER2 and FINDER using a violin plot (**Fig. 6**). Further, we applied Wilcoxon's rank-sum test and found that the TSS distances reported by FINDER were significantly less than that of BRAKER2 for *A. thaliana* and *Z. mays*. Interestingly, for *O. sativa*, BRAKER2 generated better gene models for more transcripts. Over 25% of reference gene models in *O. sativa* have no UTRs annotated which is higher compared to 15% UTR-less gene models in *A. thaliana* and *Z. mays*. This result illustrates that more FINDER transcripts have a TSS closer to the evidence as compared to the TSS of the transcripts reported by BRAKER2. This is an expected result since BRAKER2 was configured to annotate only CDS regions of transcripts. **Table 5** highlights the number of transcripts that have better agreement with the reference TSS for FINDER and BRAKER2.

Table 5 Use of RNA-Seq evidence to improve annotation of untranslated regions to aid in promoter mining and epigenetic studies

	Number of FINDER1 transcripts having TSS better than BRAKER2	Number of BRAKER2 transcripts having TSS better than FINDER1
Arabidopsis thaliana	**15063 (65%)**	8022 (35%)
Oryza sativa	**11089 (66%)**	5762 (34%)
Zea mays (NCBI)	**20721 (76%)**	6628 (24%)
Zea mays (AGPv3)	7618 (28%)	**19731 (72%)**
Zea mays (AGPv4)	**18114 (69%)**	8297 (31%)
Caenorhabditis elegans	8681 (33%)	**17730 (67%)**
Drosophila melanogaster	**10238 (63%)**	5917 (37%)
Homo sapiens	**10158 (74%)**	3486 (26%)
Hordeum vulgare	**10373 (65%)**	5607 (35%)

Fig 6.



Comparison of distance between transcription start sites of gene models predicted by BRAKER2 and FINDER. Violin plots of the distribution of the distance between the actual transcription start site (TSS) and the predicted transcription start site. In a set of well annotation complete gene structures, a higher fraction of genes is expected to have low deviation from the actual TSS. We considered genes that were reported in either BRAKER or FINDER for this analysis. Wilcoxon's rank sum test was used to compare the TSS distances between FINDER and BRAKER2. The "***" symbol implies that TSS distance for FINDER gene models was significantly less than BRAKER2 gene models.

Enhancing ground truth annotations by extending untranslated regions

Official annotations of several model organisms, used as ground truth for this study, contain transcripts with missing UTR sequences. Even though UTRs do not code for proteins, they are relevant segments of a transcript involved in several important biological processes like mRNA translation [153–155], regulation of expression [156–160]] and a number of diseases [161–165]. In the *A. thaliana* TAIR10 annotations, there are 7,888 transcripts missing either UTR; 50% of these had a rating below 2 stars.

PacBio (Menlo Park, CA) offers long-read sequencing that contain both CDS and UTRs. Therefore, we used the PacBio annotations instead of the incomplete TAIR10 transcripts to assess FINDER's performance on transcripts that were missing UTRs (Please refer to section 2.6 in Additional file 9 for more details). Out of the 7,888 TAIR10 transcripts with missing UTRs, 113 transcripts were found both in the PacBio data and the 116 short-read RNA-Seq samples. We compared the FINDER annotations against these 113 transcripts. FINDER annotations were able to recall 91.55% of the nucleotides in 113 transcripts of TAIR10 and 97.86% of PacBio transcripts. The specificity of the FINDER annotations is markedly higher with PacBio transcripts (79.67%) compared to TAIR10 transcripts (72.14%). This demonstrates that FINDER enhances and improves upon the existing annotation.

The TRITEX *H. vulgare* annotation (Morex version r2) [115], released by the International Barley Sequencing Consortium (IBSC), is devoid of UTRs. We used FINDER to update and enrich the existing annotations by flanking the CDS region with UTRs on both sides. To verify the accuracy of the gene models reported by FINDER, we used PacBio full-length mRNA sequences derived from a time course of powdery mildew infected barley leaf tissue [166, 167]. A total of 7,352 gene models from IBSC, FINDER, and PacBio had a complete intron-chain match with each other. The gene structures for more than 93% (6,886 out of 7,352) of the FINDER models were improved when compared to PacBio full-length sequences (**Additional file 7: Table**

S6). The highest F1 score achieved was 87.16. This shows that FINDER is capable of constructing accurate gene structures constituting both CDS and UTRs.

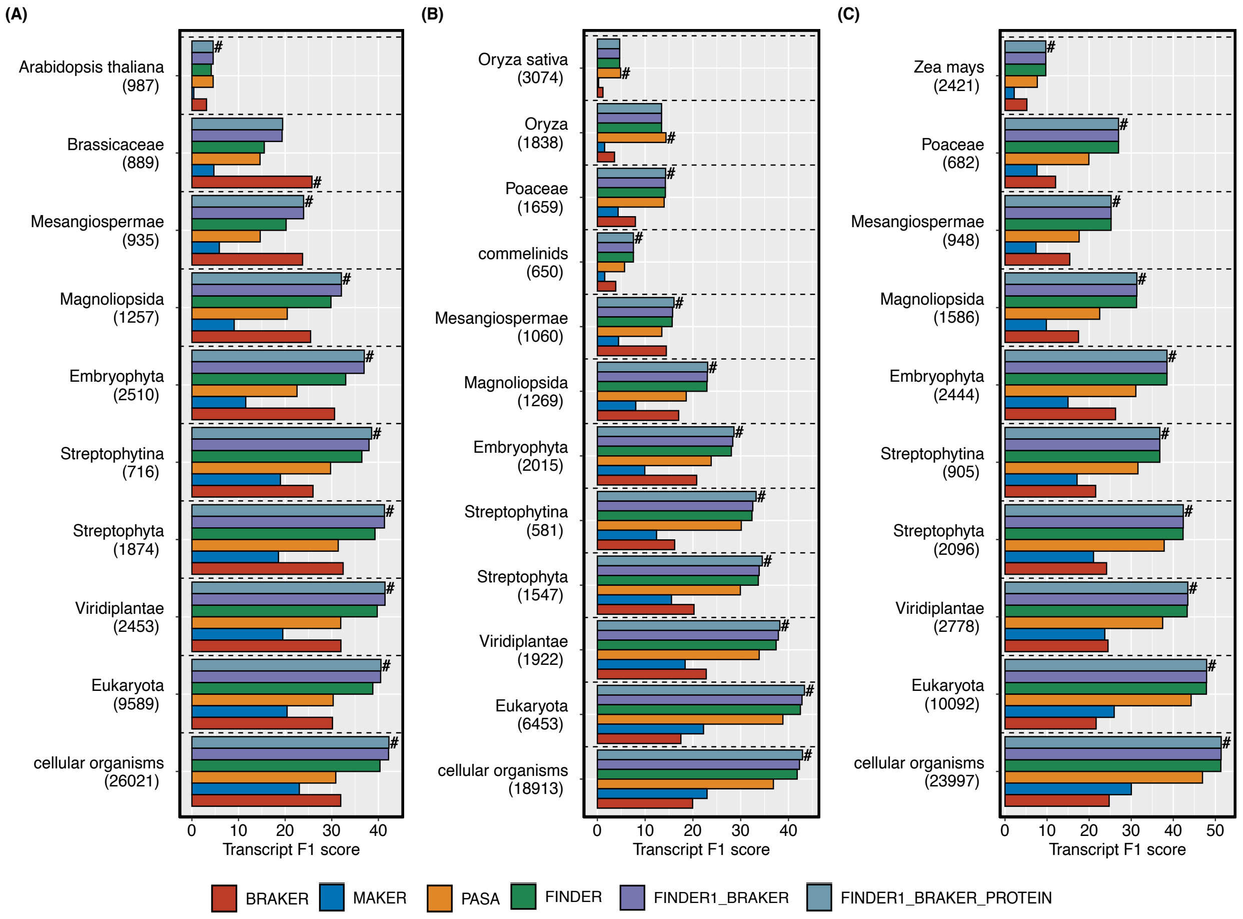
Evaluating performance with different annotations of *Zea mays*

Z. mays is an important model organism for crops and has been one of the most studied plants for genetics by researchers in several different fields [168–171] . Genes have been annotated in multiple ways using different kinds of data, resulting in substantial differences in gene structures [120]. Here we compare three alternative annotation sets of *Z. mays* – RefSeq, AGPv3, and AGPv4 and the performance of FINDER surpassed all three approaches. The transcript F1 score for FINDER gene models compared against the NCBI gene models were 43.48, whereas the F1 scores for AGPv3 and AGPv4 were 26.69 and 22.51 respectively. We observed the same trend for other annotation pipelines and reported a higher transcript F1 score for NCBI than the AGP annotations (**Table 1 and Additional file 3: Table S2).** Hence, FINDER generated high-quality gene structures with high transcript F1 scores for different *Z. mays* annotations.

Evaluating FINDER on different clades reported by Phylostratr

Genes in each organism can be categorized by their evolutionary history [172, 173]. We used Phylostratr [174] to classify genes into evolutionary strata. Here we present our results on the three model organisms – *A. thaliana, O. sativa,* and *Z. mays*. For all three, FINDER was able to accurately detect more genes in highly populated strata (**Fig. 7**). The performance of FINDER and PASA was comparable in strata with few genes. It was surprising to note that BRAKER2 was unable to identify highly conserved genes (those from the “cellular organisms” strata) since those would be easier to predict than organism specific genes. This demonstrates that FINDER is capable of effectively constructing genes from different evolutionary backgrounds.

Fig. 7



Assessment of annotation pipelines on genes from each phylostrata – Genes from three model species – (A) *Arabidopsis thaliana*, (B) *Oryza sativa* and (C) *Zea mays*, were allocated into evolutionary classes using Phylostratr. The number of genes correctly constructed by each pipeline was computed and plotted as a bar graph. Numbers below each stratum indicate the number of genes allocated to that strata. Strata having genes fewer than 500 are not shown in the graph.

FINDER constructs gene models for polyploid genomes

Being a general-purpose genome annotator, in addition to diploid organisms, FINDER can annotate the genomes of polyploid organisms. We generated gene structures of *Triticum aestivum,* a hexaploid with 120,744 annotated genes and 146,597 transcripts [115]. FINDER was able to detect 48,129 transcripts (39.9%). Out of the 130,582 transcripts predicted by FINDER, 48,104 (36.83%) matched perfectly with at least one reference annotation.

Conclusion

Identifying genes on chromosomes and deducing their structures from a plethora of evidence has been undertaken in multiple ways, with each method having advantages and disadvantages. Herein, we propose FINDER – an entirely automated, general-purpose pipeline to annotate genes in eukaryotic genomes. FINDER (1) implements an optimized mapping strategy that reduces the number of spurious mappings, (2) produces complete full-length transcripts comprising UTRs while identifying transcripts with micro-exons, (3) employs statistical CPD to modify gene boundaries and construct new genes, (4) reports more alternatively spliced transcripts as compared to other state-of-the-art annotation pipelines, and (5) assigns confidence classes to each transcript based on the evidence(s) that were used to construct those.

With a wide variety of available data for annotation, researchers often struggle to manage and optimize their usage. Several gene annotation software also offer users complicated configurations without providing substantial guidance. FINDER makes the job of gene annotation easy for bench scientists by automating the entire process from RNA-Seq data processing to gene prediction. Since FINDER does not assume the ploidy or the nucleotide composition of a genome, it could be applied to derive gene structures for a wide range of species, including non-model organisms. FINDER constructs gene models primarily from RNA-Seq data and is therefore capable of constructing tissue- and/or condition- specific isoforms which would have been impossible to obtain from ESTs only. FINDER supersedes the performance of existing software applications by utilizing read coverage information to fine-tune gene model boundaries. Instead of removing low-quality transcripts, FINDER flags them as low confidence – giving users the choice of using them as they seem fit. As a proof of concept, we provided evidence that using read coverage signal indeed enhances gene structures in a diverse set of organisms. Thus, we are confident that FINDER will pave the way for improved gene structure annotation in the future.

Availability and requirements

Project name: FINDER

Project home page: https://github.com/sagnikbanerjee15/Finder

Operating system(s): Linux, MacOS

Programming language: Python, C, C++, Perl, Shell

License: MIT

Other software requirements: All software requirements are listed in https://github.com/sagnikbanerjee15/Finder/blob/master/environment.yml

List of abbreviations

ESTs: Expressed Sequence Tags

NGS: Next Generation Sequencing

NCBI: National Center for Biotechnology Information

SRA: Sequence Read Archive

UTR: Untranslated Regions

CSV: Comma Separated Values

AED: Annotation Edit Distance

CPD: Changepoint Detection

TSS: Transcription Start Site

CDS: Coding Sequence

CPU: Central Processing Unit

cDNA: complementary DNA

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

FINDER can be accessed from https://github.com/sagnikbanerjee15/Finder

RNA-Seq samples used for annotation is included in Additional file 8: Table S7

Barley PacBio sequences have been deposited in NCBI (Project id: GSE165730)

Competing interests

The authors declare no competing interests

Funding

This research was supported by the US. Department of Agriculture, Agricultural Research Service, Project Number [5030-21000-068-00D] and [3625-21000-067-00D] through the Corn Insects and Crop Genetics Research Unit and Project Number [2030-21000-024-00D] through the Crop Improvement and Genetics Research Unit. Research supported in part by Oak Ridge Institute for Science and Education (ORISE) under U.S. Department of Energy (DOE) contract number DE-SC0014664 to SB and National Science Foundation - Plant Genome Research Program grant 13-39348 to RPW. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the USDA, ARS, DOE, ORAU/ORISE or the National Science Foundation. USDA is an equal opportunity provider and employer.

Authors' contributions

SB: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

PB: Formal Analysis, Writing – Review & Editing

MGW: Conceptualization, Resources, Supervision, Writing – Review & Editing

TZS: Conceptualization, Resources, Supervision, Writing – Review & Editing

RPW: Conceptualization, Investigation, Resources, Supervision, Writing – Review & Editing

CMA: Conceptualization, Funding Acquisition, Investigation, Project Administration, Resources, Supervision, Writing – Review & Editing

Acknowledgements

This research used resources provided by the SCINet project of the USDA Agricultural Research Service, ARS project number 0500-00093-001-00-D. The authors are also thankful to Dr. Karin Dorman (Professor, Department of Statistics, Iowa State University), for providing insightful feedbacks to compare annotations and for implementing changepoint detection. The authors thank Gregory Fuerst for taking care of submitting data to NCBI.

Additional Files

Additional file 1: Supplementary figures (S1-S9).

Additional file 2: Input to finder

Additional file 3: Annotation Edit Distance of reference transcripts as reported by each gene annotation pipeline

Additional file 4: Performance of gene annotation pipelines on coding regions of transcripts

Additional file 5: Comparison of FINDER's performance with other gene annotation pipelines on a variety of different species

Additional file 6: Comparison of different transcriptome assembly softwares on a variety of species

Additional file 7: Improvement in reference gene annotation after adding untranslated regions verified with long-read from PacBio assemblies

Additional file 8: Description of RNA-Seq data used to execute FINDER, BRAKER2, MAKER2 and PASA

Additional file 9: Supplementary text document outlining methods and some results in more details

Reference:

1. Genome List - Genome - NCBI. https://www.ncbi.nlm.nih.gov/genome/browse/#!/overview/. Accessed 12 Jan 2021.
2. Morganti S, Tarantino P, Ferraro E, D'Amico P, Viale G, Trapani D, et al. Complexity of genome sequencing and reporting: Next generation sequencing (NGS) technologies and implementation of precision medicine in real life. Critical reviews in oncology/hematology. 2019;133:171–82.
3. Koboldt DC, Steinberg KM, Larson DE, Wilson RK, Mardis ER. The next-generation sequencing revolution and its impact on genomics. Cell. 2013;155:27–38.
4. Phillips KA, Douglas MP. The Global Market for Next-Generation Sequencing Tests Continues Its Torrid Pace. The Journal of precision medicine. 2018;4.
5. Kulski JK. Next-generation sequencing—an overview of the history, tools, and "Omic" applications. Next Generation Sequencing–Advances, Applications and Challenges. 2016;:3–60.
6. Banerjee S, Mitra B, Chatterjee A, Santra A, Chatterjee B. Identification of relevant physico chemical properties of amino acids with respect to protein glycosylation prediction. In: Computing and Communication (IEMCON), 2015 International Conference and Workshop on. IEEE; 2015. p. 1–7.
7. Banerjee S, Basu S, Nasipuri M. Big Data Analytics and Its Prospects in Computational Proteomics. In: Information Systems Design and Intelligent Applications. Springer; 2015. p. 591–8.
8. Velásquez-Zapata V, Elmore JM, Banerjee S, Dorman KS, Wise RP. Y2H-SCORES: A statistical framework to infer protein-protein interactions from next-generation yeast-two-hybrid sequence data. bioarxiv. 2020. doi:https://doi.org/10.1101/2020.09.08.288365.
9. Banerjee S, Velásquez-Zapata V, Fuerst G, Elmore JM, Wise RP, Elmore M. NGPINT: a next-generation protein–protein interaction software. Briefings In Bioinformatics. 2020;2020:1–14. doi:10.1093/bib/bbaa351.
10. Rao VS, Srinivas K, Sujini GN, Kumar GN. Protein-protein interaction detection: methods and analysis. International journal of proteomics. 2014;2014.
11. Patel S, Tripathi R, Kumari V, Varadwaj P. DeepInteract: deep neural network based protein-protein interaction prediction tool. Current Bioinformatics. 2017;12:551–7.
12. Chen M, Ju CJ-T, Zhou G, Chen X, Zhang T, Chang K-W, et al. Multifaceted protein–protein interaction prediction based on siamese residual rcnn. Bioinformatics. 2019;35:i305–14.
13. Yang S, Li H, He H, Zhou Y, Zhang Z. Critical assessment and performance improvement of plant–pathogen protein–protein interaction prediction methods. Briefings in bioinformatics. 2019;20:274–87.
14. Li Y, Ilie L. SPRINT: ultrafast protein-protein interaction prediction of the entire human interactome. BMC bioinformatics. 2017;18:485.
15. Banerjee S, Ghosh D, Basu S, Nasipuri M. JUPred_MLP: Prediction of phosphorylation sites using a consensus of MLP classifiers. 2016.
16. Banerjee S, Ghosh D, Basu S, Nasipuri M. JUPred_SVM : Prediction of Phosphorylation Sites using a consensus of SVM classifiers. In: Proceedings of Fifth International Conference on Soft Computing for Problem Solving. Springer; 2016. p. 1–8.

17. Banerjee S, Nag S, Tapadar S, Ghosh S, Guha S, Bakshi S. Improving protein protein interaction prediction by choosing appropriate physiochemical properties of amino acids. In: Computing and Communication (IEMCON), 2015 International Conference and Workshop on. IEEE; 2015. p. 1–8.
18. Banerjee S, Basu S, Ghosh D, Nasipuri M. PhospredRF: Prediction of protein phosphorylation sites using a consensus of random forest classifiers. In: Computing and Communication (IEMCON), 2015 International Conference and Workshop on. IEEE; 2015. p. 1–7.
19. Luo F, Wang M, Liu Y, Zhao X-M, Li A. DeepPhos: Prediction of protein phosphorylation sites with deep learning. Bioinformatics. 2019;35:2766–73. doi:10.1093/bioinformatics/bty1051.
20. Li F, Li C, Marquez-Lago TT, Leier A, Akutsu T, Purcell AW, et al. Quokka: a comprehensive tool for rapid and accurate prediction of kinase family-specific phosphorylation sites in the human proteome. Bioinformatics. 2018;34:4223–31.
21. Song J, Wang H, Wang J, Leier A, Marquez-Lago T, Yang B, et al. PhosphoPredict: A bioinformatics tool for prediction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection. Scientific reports. 2017;7:1–19.
22. Chen H, Xue Y, Huang N, Yao X, Sun Z. MeMo: a web tool for prediction of protein methylation modifications. Nucleic acids research. 2006;34 suppl_2:W249–53.
23. Eisenhaber B, Eisenhaber F. Prediction of posttranslational modification of proteins from their amino acid sequence. In: Data Mining Techniques for the Life Sciences. Springer; 2010. p. 365–84.
24. Elmore MG, Banerjee S, Pedley KF, Ruck A, Whitham SA. De novo transcriptome of Phakopsora pachyrhizi uncovers putative effector repertoire during infection. Physiological and Molecular Plant Pathology. 2020;110.
25. Frantzeskakis L, Kracher B, Kusch S, Yoshikawa-Maekawa M, Bauer S, Pedersen C, et al. Signatures of host specialization and a recent transposable element burst in the dynamic one-speed genome of the fungal barley powdery mildew pathogen. BMC Genomics. 2018;19:381. doi:10.1186/s12864-018-4750-6.
26. Sperschneider J. Machine learning in plant–pathogen interactions: empowering biological predictions from field scale to genome scale. New Phytologist. 2019;:nph.15771. doi:10.1111/nph.15771.
27. Sperschneider J, Dodds PN, Singh KB, Taylor JM. ApoplastP: prediction of effectors and plant proteins in the apoplast using machine learning. New Phytologist. 2017.
28. Sperschneider J, Gardiner DM, Dodds PN, Tini F, Covarelli L, Singh KB, et al. EffectorP: predicting fungal effector proteins from secretomes using machine learning. The New Phytologist. 2016;210:743–61. doi:10.1111/nph.13794.
29. Magnan CN, Baldi P. SSpro/ACCpro 5: Almost Perfect Prediction of Protein Secondary Structure and Relative Solvent Accessibility Using Profiles, Machine Learning, and Structural Similarity. Bioinformatics. 2014;:btu352.
30. McGuffin LJ, Bryson K, Jones DT. The PSIPRED protein structure prediction server. Bioinformatics (Oxford, England). 2000;16:404–5.
31. Laskowski RA, Watson JD, Thornton JM. Protein function prediction using local 3D templates. Journal of molecular biology. 2005;351:614–26. doi:10.1016/j.jmb.2005.05.067.
32. Banerjee S, Guha S, Dutta A, Dutta S. Improvement of protein disorder prediction by brainstorming consensus. In: Computing and Communication (IEMCON), 2015 International Conference and Workshop on. IEEE; 2015. p. 1–7.

33. Salzberg SL. Next-generation genome annotation: we still struggle to get it right. BioMed Central; 2019. doi:10.1186/s13059-019-1715-2.
34. del Angel VD, Hjerde E, Sterck L, Capella-Gutierrez S, Notredame C, Pettersson OV, et al. Ten steps to get started in Genome Assembly and Annotation. F1000Research. 2018;7.
35. Richards S. Full disclosure: Genome assembly is still hard. PLoS biology. 2018;16:e2005894.
36. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith Jr RK, Hannick LI, et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic acids research. 2003;31:5654–66.
37. Salamov A, Solovyev V. Fgenesh multiple gene prediction program. 1998.
38. Solovyev V, Kosarev P, Seledsov I, Vorobyev D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome biology. 2006;7:S10.
39. Kleffe J, Hermann K, Vahrson W, Wittig B, Brendel V. GeneGenerator--a flexible algorithm for gene prediction and its application to maize sequences. Bioinformatics (Oxford, England). 1998;14:232–43.
40. Schweikert G, Zien A, Zeller G, Behr J, Dieterich C, Ong CS, et al. mGene: accurate SVM-based gene finding with an application to nematode genomes. Genome research. 2009;19:2133–43.
41. Schlueter SD, Dong Q, Brendel V. GeneSeqer@ PlantGDB: gene structure prediction in plant genomes. Nucleic Acids Research. 2003;31:3597–600.
42. Cantarel BL, Korf I, Robb SMCC, Parra G, Ross E, Moore B, et al. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research. 2008;18:188–96. doi:10.1101/gr.6743907.
43. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC bioinformatics. 2011;12:491.
44. Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, et al. MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant physiology. 2014;164:513–24.
45. Campbell MS, Holt C, Moore B, Yandell M. Genome annotation and curation using MAKER and MAKER-P. Current Protocols in Bioinformatics. 2014;48:4–11. doi:10.1002/0471250953.bi0411s48.
46. Vonk FJ, Casewell NR, Henkel C v, Heimberg AM, Jansen HJ, McCleary RJR, et al. The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. Proceedings of the National Academy of Sciences. 2013;110:20651–6.
47. Keane M, Semeiks J, Webb AE, Li YI, Quesada V, Craig T, et al. Insights into the evolution of longevity from the bowhead whale genome. Cell reports. 2015;10:112–22.
48. Zhang J, Fu X-X, Li R-Q, Zhao X, Liu Y, Li M-H, et al. The hornwort genome and early land plant evolution. Nature plants. 2020;6:107–18.
49. Gray MW, Burger G, Derelle R, Klimeš V, Leger MM, Sarrasin M, et al. The draft nuclear genome sequence and predicted mitochondrial proteome of Andalucia godoyi, a protist with the most gene-rich and bacteria-like mitochondrial genome. BMC biology. 2020;18:1–35.
50. Peng C, Ren J-L, Deng C, Jiang D, Wang J, Qu J, et al. The genome of Shaw’s sea snake (Hydrophis curtus) reveals secondary adaptation to its marine environment. Molecular Biology and Evolution. 2020.

51. Weitemier K, Straub SCK, Fishbein M, Bailey CD, Cronn RC, Liston A. A draft genome and transcriptome of common milkweed (Asclepias syriaca) as resources for evolutionary, ecological, and molecular studies in milkweeds and Apocynaceae. PeerJ. 2019;7:e7649.
52. Zhang J, Zhang X, Tang H, Zhang Q, Hua X, Ma X, et al. Allele-defined genome of the autopolyploid sugarcane Saccharum spontaneum L. Nature genetics. 2018;50:1565–73.
53. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research. 2008;18:821–9. https://www.ebi.ac.uk/~zerbino/velvet/.
54. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature biotechnology. 2011;29:644.
55. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature biotechnology. 2011;29:644–52.
56. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology. 2012;19:455–77.
57. Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, et al. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics. 2014;30:1660–6.
58. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature biotechnology. 2010;28:511.
59. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature Biotechnology. 2015;33:290–5. doi:10.1038/nbt.3122.
60. Liu R, Dickerson J. Strawberry: Fast and accurate genome-guided transcript reconstruction and quantification from RNA-Seq. PLoS computational biology. 2017;13:e1005851.
61. Shao M, Kingsford C. Accurate assembly of transcripts through phase-preserving graph decomposition. Nature Biotechnology. 2017;35:1167–9. doi:10.1038/nbt.4020.
62. Kovaka S, Zimin A v, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. Genome Biology. 2019;20:1–13.
63. Song L, Sabunciyan S, Yang G, Florea L. A multi-sample approach increases the accuracy of transcript assembly. Nature Communications. 2019;10:5000. doi:10.1038/s41467-019-12990-0.
64. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS: Table 1. Bioinformatics. 2016;32:767–9. doi:10.1093/bioinformatics/btv661.
65. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-genome annotation with BRAKER. In: Gene Prediction. Springer; 2019. p. 65–95.
66. Steijger T, Abril JF, Engström PG, Kokocinski F, Akerman M, Alioto T, et al. Assessment of transcript reconstruction methods for RNA-seq. Nature methods. 2013;10:1177–84.
67. Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic acids research. 2014;42:e119–e119. doi:10.1093/nar/gku557.
68. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24:637–44.
69. Korf I. Gene finding in novel genomes. BMC bioinformatics. 2004;5:59.

70. Leinonen R, Sugawara H, Shumway M, Collaboration INSD. The sequence read archive. Nucleic acids research. 2010;39 suppl_1:D19–21.
71. Dobin A, Gingeras TR, Spring C, Flores R, Sampson J, Knight R, et al. Mapping RNA-seq with STAR. Curr Protoc Bioinformatics. 2016;51:586–97.
72. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.
73. Tang S, Lomsadze A, Borodovsky M. Identification of protein coding regions in RNA transcripts. Nucleic Acids Research. 2015;43:e78. doi:10.1093/nar/gkv227.
74. Goodstadt L. Ruffus: a lightweight Python library for computational pipelines. Bioinformatics. 2010;26:2778–9.
75. Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Alioto T, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. Nature methods. 2013;10:1185–91.
76. Rapazote-Flores P, Bayer M, Milne L, Mayer C-D, Fuller J, Guo W, et al. BaRTv1. 0: an improved barley reference transcript dataset to determine accurate changes in the barley transcriptome using RNA-seq. BMC genomics. 2019;20:1–17.
77. Ustianenko D, Weyn-Vanhentenryck SM, Zhang C. Microexons: discovery, regulation, and function. Wiley Interdisciplinary Reviews: RNA. 2017;8:e1418.
78. Curry-Hyde A, Chen BJ, Mills JD, Janitz M. Microexons: novel regulators of the transcriptome. Journal of Human Transcriptome. 2018;2:1–6.
79. Wen F, Li F, Xia H, Lu X, Zhang X, Li Y. The impact of very short alternative splicing on protein structures and functions in the human genome. Trends in Genetics. 2004;20:232–6.
80. Sakharkar MK, Chow VTK, Kangueane P. Distributions of exons and introns in the human genome. In silico biology. 2004;4:387–93.
81. Mano F, Aoyanagi T, Kozaki A. Atypical splicing accompanied by skipping conserved micro-exons produces unique WRINKLED1, an AP2 domain transcription factor in rice plants. Plants. 2019;8:207.
82. Song Q, Lv F, Tahir ul Qamar M, Xing F, Zhou R, Li H, et al. Identification and analysis of micro-exon genes in the rice genome. International journal of molecular sciences. 2019;20:2685.
83. Bulman S, Ridgway HJ, Eady C, Conner AJ. Intron-rich gene structure in the intracellular plant parasite Plasmodiophora brassicae. Protist. 2007;158:423–33.
84. Wang X. Protein and Proteome Atlas for Plants under Stresses: New Highlights and Ways for Integrated Omics in Post-Genomics Era. 2019.
85. Guo L, Liu C-M. A single-nucleotide exon found in Arabidopsis. Scientific reports. 2015;5:18087.
86. Gonatopoulos-Pournatzis T, Wu M, Braunschweig U, Roth J, Han H, Best AJ, et al. Genome-wide CRISPR-Cas9 interrogation of splicing networks reveals a mechanism for recognition of autism-misregulated neuronal microexons. Molecular cell. 2018;72:510–24.
87. Consortium Gte. Human genomics. The human transcriptome across tissues and individuals. Science. 2015;348:660–5.
88. Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, Babor M, et al. A highly conserved program of neuronal microexons is misregulated in autistic brains. Cell. 2014;159:1511–23.
89. Torres-Méndez A, Bonnal S, Marquez Y, Roth J, Iglesias M, Permanyer J, et al. A novel protein domain in an ancestral splicing factor drove the evolution of neural microexons. Nature Ecology & Evolution. 2019;3:691–701.

90. Parras A, Anta H, Santos-Galindo M, Swarup V, Elorza A, Nieto-González JL, et al. Autism-like phenotype and risk gene mRNA deadenylation by CPEB4 mis-splicing. Nature. 2018;560:441–6.

91. Wu J, Anczukow O, Krainer AR, Zhang MQ, Zhang C. OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. Nucleic acids research. 2013;41:5149–63.

92. Kawahara Y, Sugiyama M. Change-point detection in time-series data by direct density-ratio estimation. In: Proceedings of the 2009 SIAM International Conference on Data Mining. SIAM; 2009. p. 389–400.

93. Lund R, Wang XL, Lu QQ, Reeves J, Gallagher C, Feng Y. Changepoint detection in periodic and autocorrelated time series. Journal of Climate. 2007;20:5178–90.

94. Kawahara Y, Yairi T, Machida K. Change-point detection in time-series data based on subspace identification. In: Seventh IEEE International Conference on Data Mining (ICDM 2007). IEEE; 2007. p. 559–64.

95. Takeuchi J, Yamanishi K. A unifying framework for detecting outliers and change points from time series. IEEE transactions on Knowledge and Data Engineering. 2006;18:482–92.

96. Aalvik Stranden S. A Supervised Sliding Window Approach for Change Point Detection in Multivariate Time Series. 2020.

97. Tartakovsky AG, Rozovskii BL, Blazek RB, Kim H. A novel approach to detection of intrusions in computer networks via adaptive sequential and batch-sequential change-point detection methods. IEEE transactions on signal processing. 2006;54:3372–82.

98. Klanderman MC, Newhart KB, Cath TY, Hering AS. Fault isolation for a complex decentralized waste water treatment facility. Journal of the Royal Statistical Society Series C. 2020;69:931–51.

99. Quinlan AR. BEDTools: the Swiss-army tool for genome feature analysis. Current protocols in bioinformatics. 2014;47:11–2.

100. Killick R, Eckley I. changepoint: An R package for changepoint analysis. Journal of statistical software. 2014;58:1–19.

101. Xiang S, Huang Z, Wang T, Han Z, Christina YY, Ni D, et al. Condition-specific gene co-expression network mining identifies key pathways and regulators in the brain tissue of Alzheimer's disease patients. BMC medical genomics. 2018;11:115.

102. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology. 1990;215:403–10.

103. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. BMC bioinformatics. 2005;6:31.

104. Cheng C, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Araport11: a complete reannotation of the Arabidopsis thaliana reference genome. The Plant Journal. 2017;89:789–804.

105. Li J-Y, Wang J, Zeigler RS. The 3,000 rice genomes project: new opportunities and challenges for future rice research. Gigascience. 2014;3:2047-217X.

106. Duitama J, Silva A, Sanabria Y, Cruz DF, Quintero C, Ballen C, et al. Whole genome sequencing of elite rice cultivars as a comprehensive information resource for marker assisted selection. PLoS One. 2015;10:e0124617.

107. Zhao K, Wright M, Kimball J, Eizenga G, McClung A, Kovach M, et al. Genomic diversity and introgression in O. sativa reveal the impact of domestication and breeding on the rice genome. PloS one. 2010;5:e10780.

108. Wang B, Tseng E, Regulski M, Clark TA, Hon T, Jiao Y, et al. Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nature Communications. 2016;7:11708. doi:10.1038/ncomms11708.
109. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. science. 2009;326:1112–5.
110. The_C_elegans_Sequencing_Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science. 1998;282:2012–8.
111. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome research. 2008;18:1979–90. doi:10.1101/gr.081612.108.
112. Drosophila_consortium. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 2007;450:203.
113. International_Human_Genome_Sequencing_consortium. Initial sequencing and analysis of the human genome. Nature. 2001;409:860–921.
114. Hood L, Rowen L. The human genome project: big science transforms biology and medicine. Genome medicine. 2013;5:79.
115. Monat C, Padmarasu S, Lux T, Wicker T, Gundlach H, Himmelbach A, et al. TRITEX: chromosome-scale sequence assembly of Triticeae genomes with open-source tools. Genome biology. 2019;20:284.
116. Appels R, Eversole K, Stein N, Feuillet C, Keller B, Rogers J, et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science. 2018;361.
117. Krasileva K v, Vasquez-Gross HA, Howell T, Bailey P, Paraiso F, Clissold L, et al. Uncovering hidden variation in polyploid wheat. Proceedings of the National Academy of Sciences. 2017;114:E913–21.
118. Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, Wright J, et al. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Genome research. 2017;27:885–96.
119. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, et al. RefSeq: an update on mammalian reference sequences. Nucleic acids research. 2014;42:D756–63.
120. Tello-Ruiz MK, Naithani S, Stein JC, Gupta P, Campbell M, Olson A, et al. Gramene 2018: unifying comparative genomics and pathway resources for plant research. Nucleic acids research. 2018;46:D1181–9.
121. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, et al. Improved maize reference genome with single-molecule technologies. Nature. 2017;546:524–7.
122. Eilbeck K, Moore B, Holt C, Yandell M. Quantitative measures for the management and comparison of annotated genomes. BMC bioinformatics. 2009;10:67.
123. Venturini L, Caim S, Kaithakottil GG, Mapleson DL, Swarbreck D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. GigaScience. 2018;7. doi:10.1093/gigascience/giy093.
124. Bruna T, Hoff K, Stanke M, Lomsadze A, Borodovsky M. BRAKER2: Automatic Eukaryotic Genome Annotation with GeneMark-EP+ and AUGUSTUS Supported by a Protein Database. bioRxiv. 2020.
125. TAIR. Documentation for the TAIR gene model and exon confidence ranking system. 2009. http://plantta.jcvi.org/. Accessed 9 Oct 2020.

126. Sreenivasamurthy SK, Madugundu AK, Patil AH, Dey G, Mohanty AK, Kumar M, et al. Mosquito-borne diseases and Omics: tissue-restricted expression and alternative splicing revealed by transcriptome profiling of Anopheles stephensi. Omics: a journal of integrative biology. 2017;21:488–97.
127. Azlan A, Obeidat SM, Yunus MA, Azzam G. Transcriptome profiles and novel lncRNA identification of Aedes aegypti cells in response to dengue virus serotype 1. BioRxiv. 2018;:422170.
128. Azlan A, Halim MA, Azzam G. Genome-wide identification and characterization of long intergenic noncoding RNAs in the regenerative flatworm Macrostomum lignano. Genomics. 2020;112:1273–81.
129. Qi S, Akter S, Li S. Identification of Novel lincRNA and Co-Expression Network Analysis Using RNA-Sequencing Data in Plants. In: Plant Long Non-Coding RNAs. Springer; 2019. p. 207–21.
130. Beisel NS, Noble J, Barbazuk WB, Paul A-L, Ferl RJ. Spaceflight-induced alternative splicing during seedling development in Arabidopsis thaliana. NPJ microgravity. 2019;5:1–5.
131. Wang C, Wallerman O, Arendt M-L, Sundstrom E, Karlsson A, Nordin J, et al. A new long-read dog assembly uncovers thousands of exons and functional elements missing in the previous reference. bioRxiv. 2020.
132. Liu S, Aagaard A, Bechsgaard J, Bilde T. DNA methylation patterns in the social spider, Stegodyphus dumicola. Genes. 2019;10:137.
133. Wu S, Gao S, Wang S, Meng J, Wickham J, Luo S, et al. A Reference Genome of Bursaphelenchus mucronatus Provides New Resources for Revealing Its Displacement by Pinewood Nematode. Genes. 2020;11:570.
134. Wang P, Luo Y, Huang J, Gao S, Zhu G, Dang Z, et al. The genome evolution and domestication of tropical fruit mango. Genome biology. 2020;21:1–17.
135. Cieślik M, Chinnaiyan AM. Cancer transcriptome profiling at the juncture of clinical translation. Nature Reviews Genetics. 2018;19:93.
136. Lorenzi L, Avila Cobos F, Decock A, Everaert C, Helsmoortel H, Lefever S, et al. Long noncoding RNA expression profiling in cancer: challenges and opportunities. Genes, Chromosomes and Cancer. 2019;58:191–9.
137. Yang J, Moeinzadeh M-H, Kuhl H, Helmuth J, Xiao P, Haas S, et al. Haplotype-resolved sweet potato genome traces back its hexaploidization history. Nature plants. 2017;3:696–703.
138. Sun Z, Nair A, Chen X, Prodduturi N, Wang J, Kocher J-P. UClncR: Ultrafast and comprehensive long non-coding RNA detection from RNA-seq. Scientific reports. 2017;7:1–10.
139. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature protocols. 2012;7:562–78.
140. Arrigoni A, Ranzani V, Rossetti G, Panzeri I, Abrignani S, Bonnal RJP, et al. Analysis RNA-seq and Noncoding RNA. In: Polycomb Group Proteins. Springer; 2016. p. 125–35.
141. Ghosh S, Chan C-KK. Analysis of RNA-Seq data using TopHat and Cufflinks. In: Plant Bioinformatics. Springer; 2016. p. 339–61.
142. Qi X, Xie S, Liu Y, Yi F, Yu J. Genome-wide annotation of genes and noncoding RNAs of foxtail millet in response to simulated drought stress by deep sequencing. Plant Molecular Biology. 2013;83:459–73.
143. Marchant A, Mougel F, Mendonça V, Quartier M, Jacquin-Joly E, da Rosa JA, et al. Comparing de novo and reference-based transcriptome assembly strategies by applying them to

the blood-sucking bug Rhodnius prolixus. Insect biochemistry and molecular biology. 2016;69:25–33.

144. Li A, Zhang J, Zhou Z, Wang L, Liu Y, Liu Y. ALDB: a domestic-animal long noncoding RNA database. PloS one. 2015;10:e0124003.

145. Cooper SJ, Trinklein ND, Anton ED, Nguyen L, Myers RM. Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome. Genome research. 2006;16:1–10.

146. Brown RH, Gross SS, Brent MR. Begin at the beginning: Predicting genes with 5′ UTRs. Genome research. 2005;15:742–7.

147. Ohler U, Liao G, Niemann H, Rubin GM. Computational analysis of core promoters in the Drosophila genome. Genome biology. 2002;3:research0087-1.

148. Batut P, Gingeras TR. RAMPAGE: Promoter Activity Profiling by Paired-End Sequencing of 5′-Complete cDNAs. Current protocols in molecular biology. 2013;104:25B – 11.

149. Adiconis X, Haber AL, Simmons SK, Levy Moonshine A, Ji Z, Busby MA, et al. Comprehensive comparative analysis of 5′-end RNA-sequencing methods. Nature methods. 2018;15:505–11.

150. Shiraki T, Kondo S, Katayama S, Waki K, Kasukawa T, Kawaji H, et al. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proceedings of the National Academy of Sciences. 2003;100:15776–81.

151. Holmqvist E, Wright PR, Li L, Bischler T, Barquist L, Reinhardt R, et al. Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking in vivo. The EMBO journal. 2016;35:991–1011.

152. Hickman R, van Verk MC, van Dijken AJH, Mendes MP, Vroegop-Vos IA, Caarls L, et al. Architecture and dynamics of the jasmonic acid gene regulatory network. The Plant Cell Online. 2017;:tpc-00958.

153. Jackson RJ, Standart N. Do the poly (A) tail and 3′ untranslated region control mRNA translation? Cell. 1990;62:15–24.

154. Meijer HA, Thomas AAM. Control of eukaryotic protein synthesis by upstream open reading frames in the 5′-untranslated region of an mRNA. Biochemical Journal. 2002;367:1–11.

155. Miller GM, Madras BK. Polymorphisms in the 3′-untranslated region of human and monkey dopamine transporter genes affect reporter gene expression. Molecular psychiatry. 2002;7:44–55.

156. Wu S, Huang S, Ding J, Zhao Y, Liang L, Liu T, et al. Multiple microRNAs modulate p21Cip1/Waf1 expression by directly targeting its 3′ untranslated region. Oncogene. 2010;29:2302–8.

157. Dixon DA, Kaplan CD, McIntyre TM, Zimmerman GA, Prescott SM. Post-transcriptional control of cyclooxygenase-2 gene expression The role of the 3′-untranslated region. Journal of Biological Chemistry. 2000;275:11750–7.

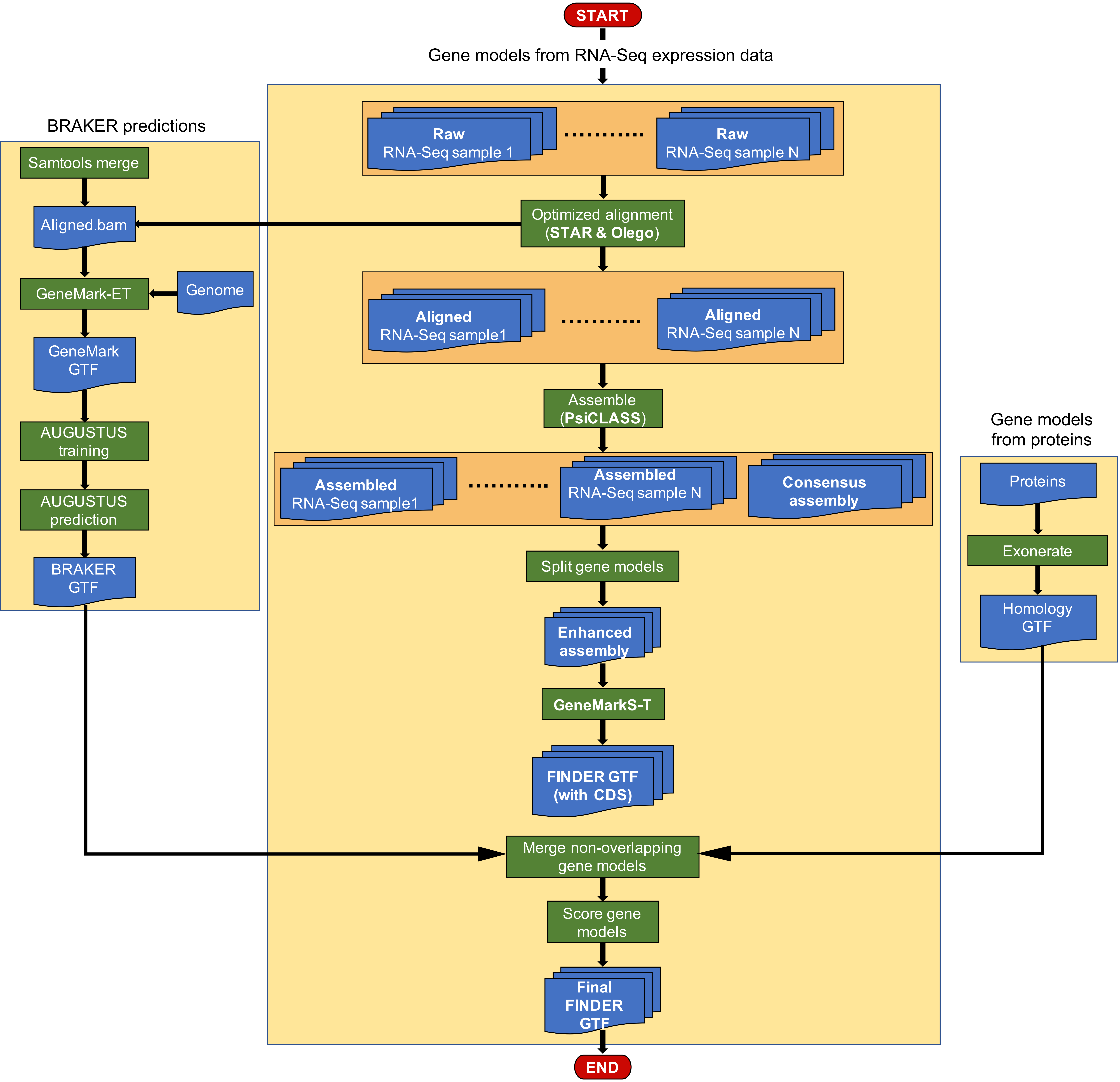
158. Gu S, Jin L, Zhang F, Sarnow P, Kay MA. Biological basis for restriction of microRNA targets to the 3′ untranslated region in mammalian mRNAs. Nature structural & molecular biology. 2009;16:144.

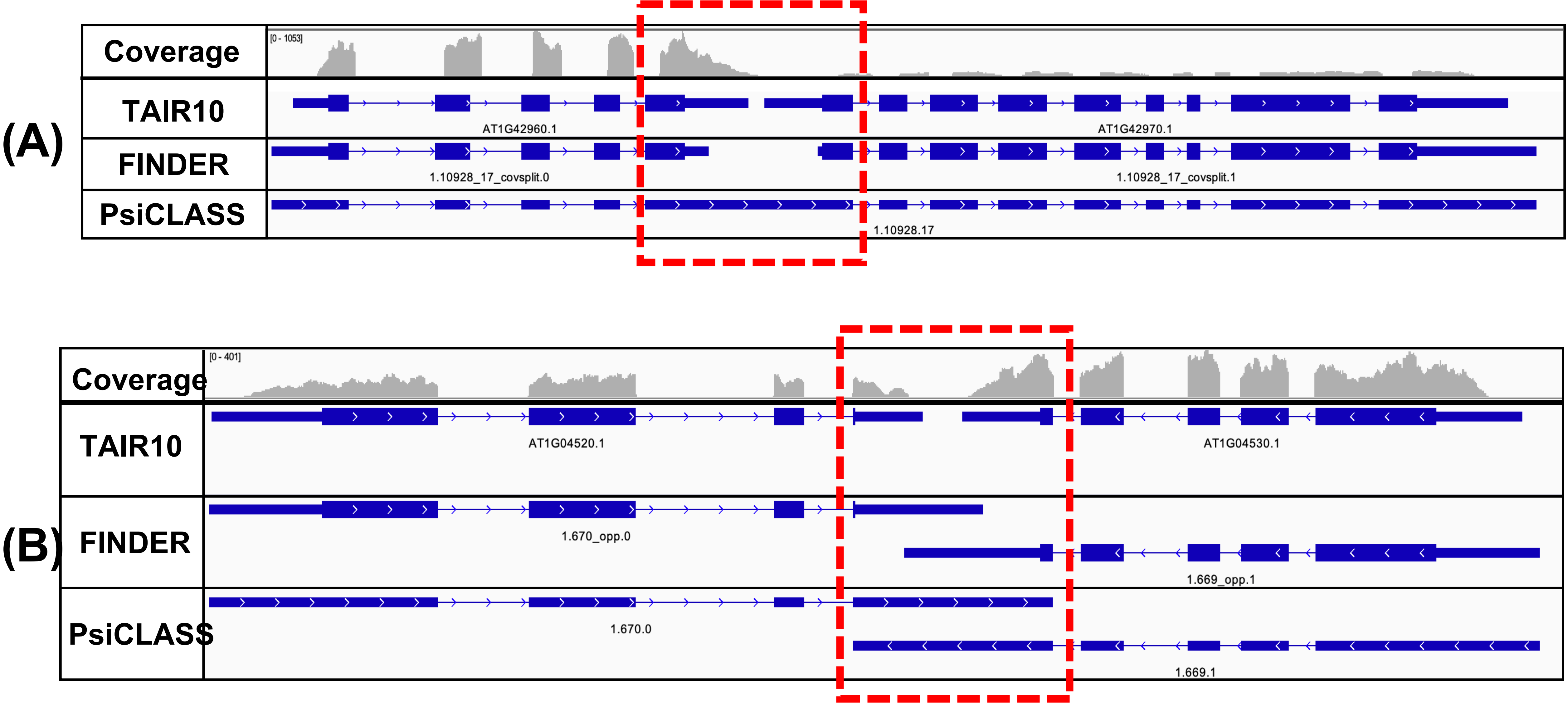
159. Eberle AB, Stalder L, Mathys H, Orozco RZ, Mühlemann O. Posttranscriptional gene regulation by spatial rearrangement of the 3′ untranslated region. PLoS Biol. 2008;6:e92.

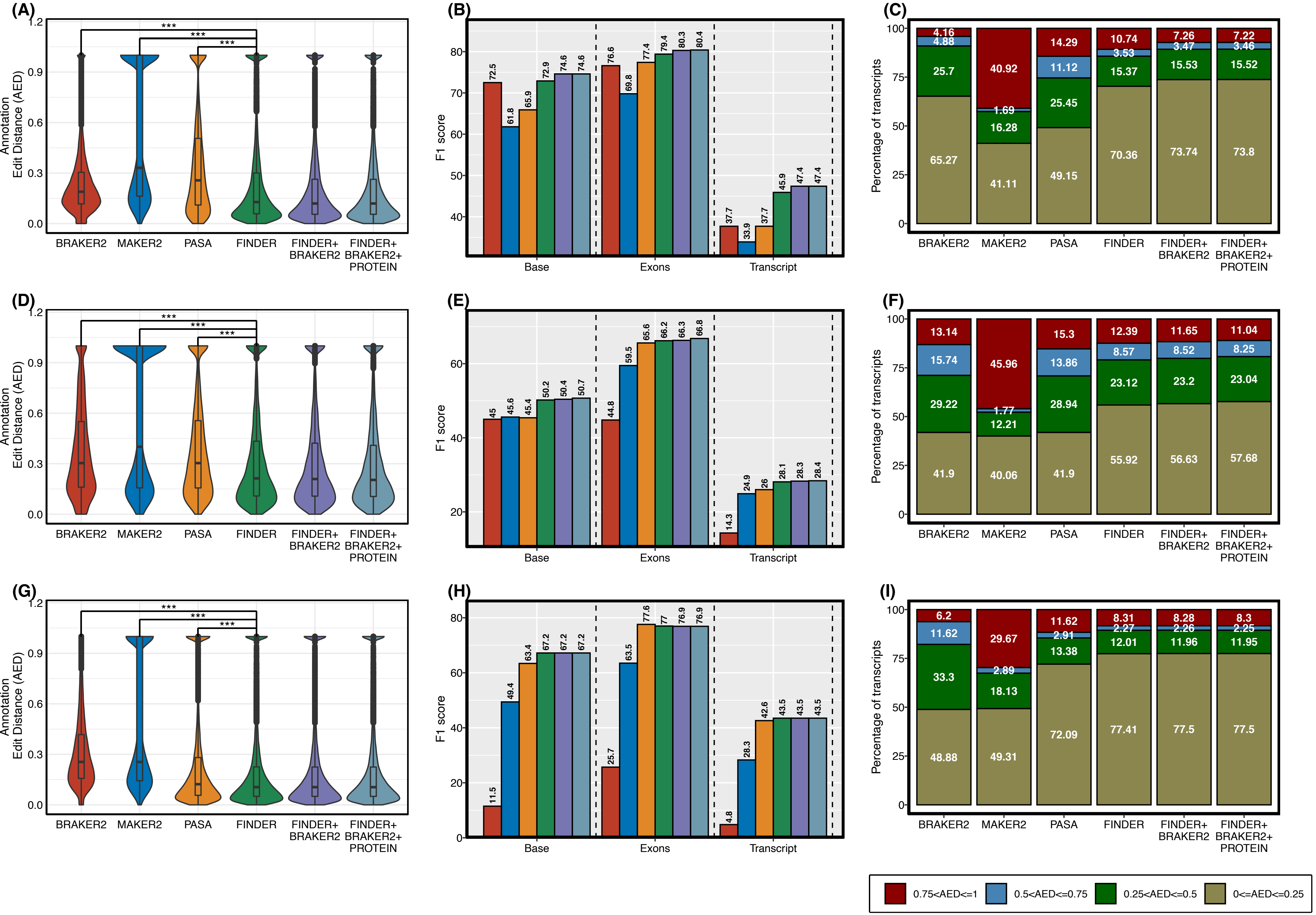
160. Halterman DA, Wise RP. Upstream open reading frames of the barley Mla13 powdery mildew resistance gene function co-operatively to down-regulate translation. Molecular plant pathology. 2006;7:167–76.

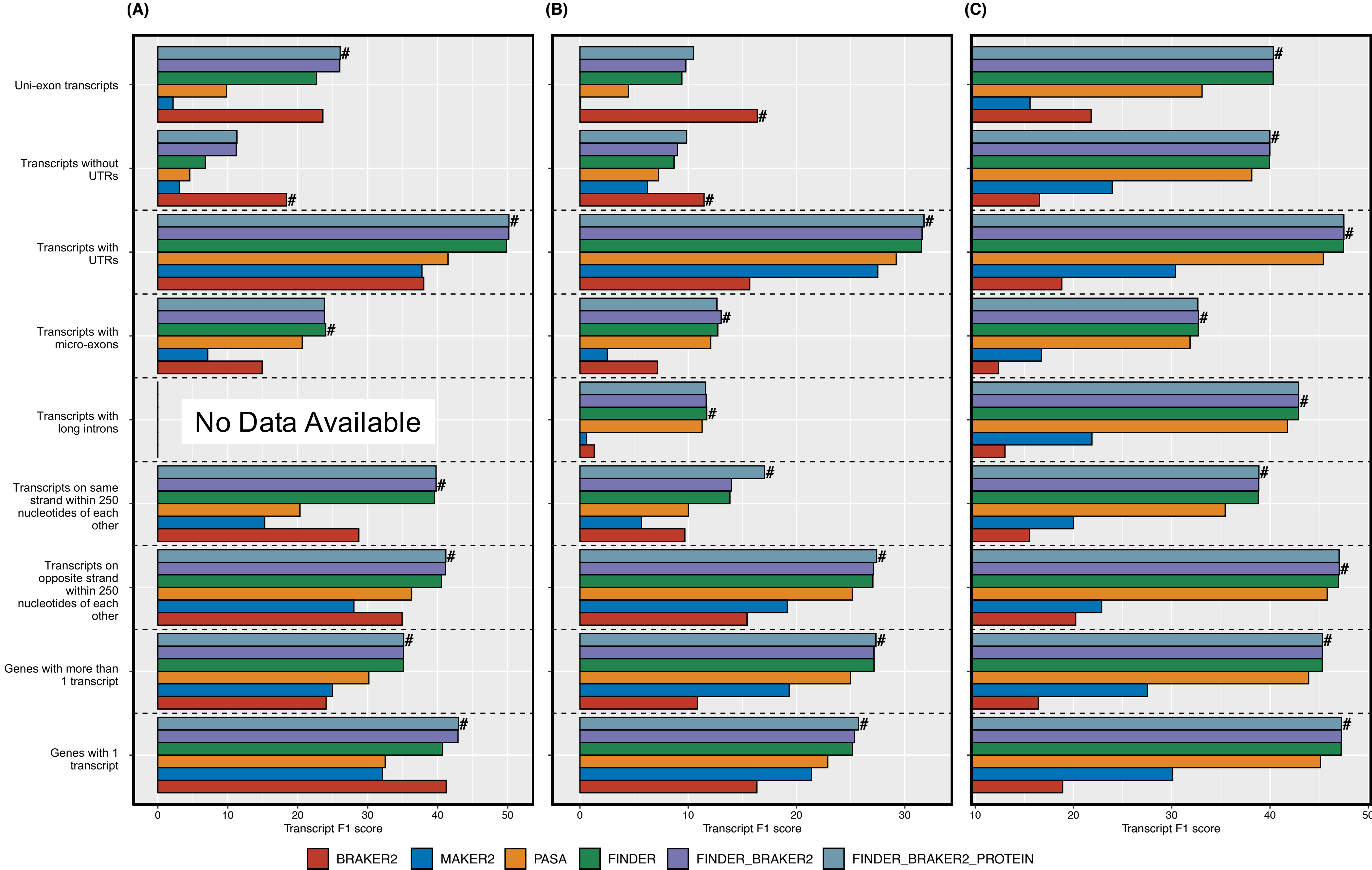
161. Awata T, Inoue K, Kurihara S, Ohkubo T, Watanabe M, Inukai K, et al. A common polymorphism in the 5′-untranslated region of the VEGF gene is associated with diabetic retinopathy in type 2 diabetes. Diabetes. 2002;51:1635–9.
162. Rogers JT, Randall JD, Cahill CM, Eder PS, Huang X, Gunshin H, et al. An iron-responsive element type II in the 5′-untranslated region of the Alzheimer's amyloid precursor protein transcript. Journal of Biological Chemistry. 2002;277:45518–28.
163. Chin LJ, Ratner E, Leng S, Zhai R, Nallur S, Babar I, et al. A SNP in a let-7 microRNA complementary site in the KRAS 3′ untranslated region increases non–small cell lung cancer risk. Cancer research. 2008;68:8535–40.
164. Halterman DA, Wei F, Wise RP. Powdery mildew-induced Mla mRNAs are alternatively spliced and contain multiple upstream open reading frames. Plant physiology. 2003;131:558–67. doi:10.1104/pp.014407.
165. Conne B, Stutz A, Vassalli J-D. The 3′ untranslated region of messenger RNA: a molecular 'hotspot'for pathology? Nature medicine. 2000;6:637–41.
166. Hunt M, Banerjee S, Surana P, Liu M, Fuerst G, Mathioni S, et al. Small RNA discovery in the interaction between barley and the powdery mildew pathogen. BMC genomics. 2019;20:610.
167. Chapman AVE, Matthew H, Surana P, Velásquez-Zapata V, Xu W, Fuerst G, et al. Disruption of barley immunity to powdery mildew by an in-frame Lys-Leu deletion in the essential protein SGT1. Oxford Genetics. 2020.
168. Dai X, Xu Z, Liang Z, Tu X, Zhong S, Schnable JC. Non-Homology-Based Prediction of Gene Functions. 2019;:1–18.
169. Duvick DN. The contribution of breeding to yield advances in maize (Zea mays L.). Advances in agronomy. 2005;86:83–145.
170. Agrama HAS, Moussa ME. Mapping QTLs in breeding for drought tolerance in maize (Zea mays L.). euphytica. 1996;91:89–97.
171. Maazou A-RS, Tu J, Qiu J, Liu Z. Breeding for drought tolerance in maize (Zea mays L.). American Journal of Plant Sciences. 2016;7:1858.
172. Bhandary P, Seetharam AS, Arendsee ZW, Hur M, Wurtele ES. Raising orphans from a metadata morass: A researcher's guide to re-use of public 'omics data. 2018. doi:10.1016/j.plantsci.2017.10.014.
173. Arendsee ZW, Li L, Wurtele ES. Coming of age: orphan genes in plants. Trends in Plant Science. 2014;19:698–708. doi:10.1016/J.TPLANTS.2014.07.003.
174. Arendsee Z, Li J, Singh U, Seetharam A, Dorman K, Wurtele ES. phylostratr: A framework for phylostratigraphy. Bioinformatics. 2019;35:3617–27.

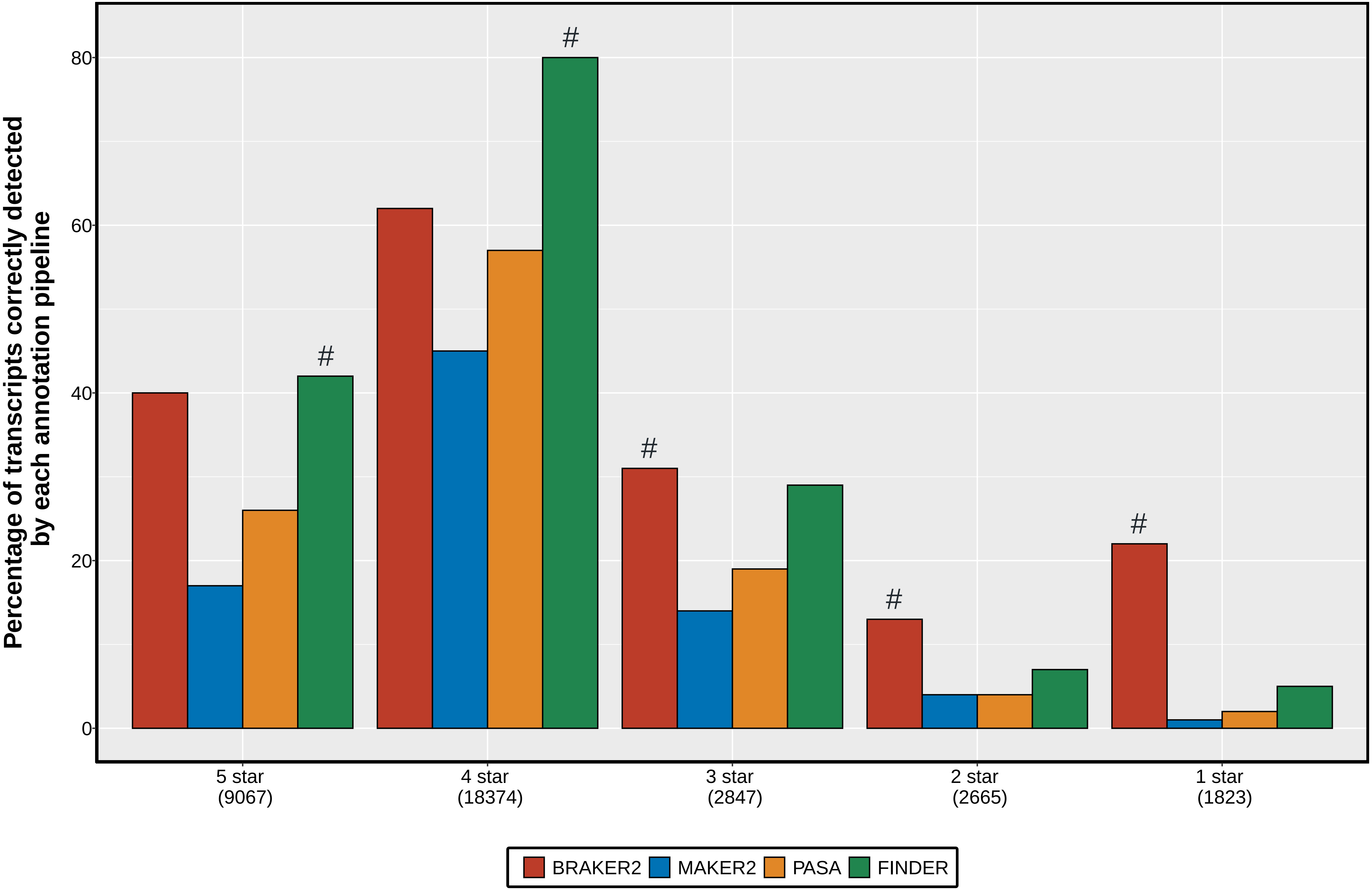
16th Aug, 2020

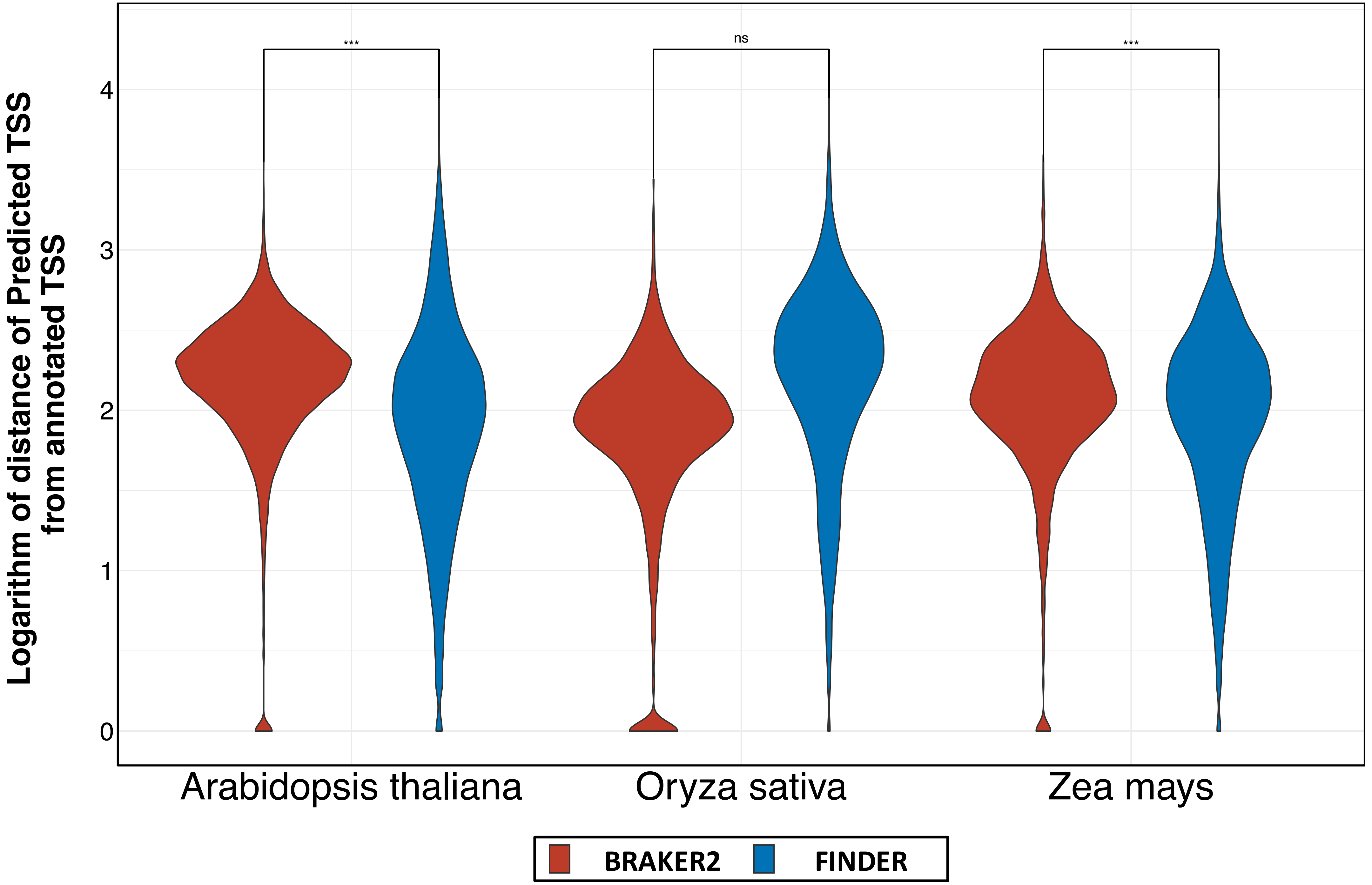



(A)
Coverage
[0 - 1053]
TAIR10
AT1G42960.1
AT1G42970.1
FINDER
1.10928_17_covsplit.0
1.10928_17_covsplit.1
PsiCLASS
1.10928.17
(B)
Coverage
[0 - 401]
TAIR10
AT1G04520.1
AT1G04530.1
FINDER
1.670_opp.0
1.669_opp.1
PsiCLASS
1.670.0
1.669.1

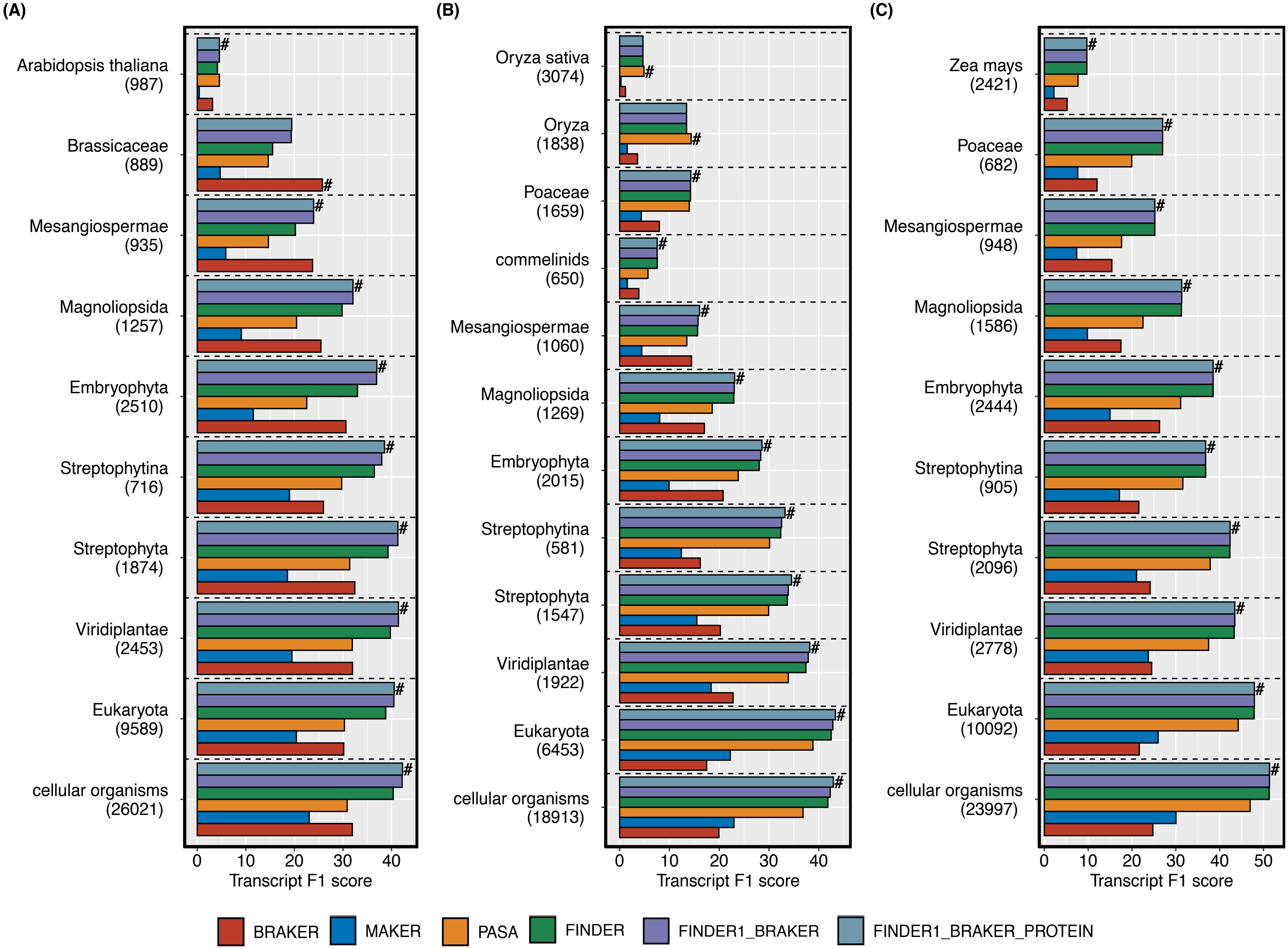

(A)
(B)
(C)
(D)
(E)
(F)
(G)
(H)
(I)
Annotation
Edit Distance (AED)
0.0
0.3
0.6
0.9
1.2
***
***
***
BRAKER2
MAKER2
PASA
FINDER
FINDER+
BRAKER2
FINDER+
BRAKER2+
PROTEIN
F1 score
Base
Exons
Transcript
72.5
61.8
65.9
72.9
74.6
74.6
76.6
69.8
77.4
79.4
80.3
80.4
37.7
33.9
37.7
45.9
47.4
47.4
Percentage of transcripts
0
25
50
75
100
4.16
4.88
25.7
65.27
40.92
1.69
16.28
41.11
14.29
11.12
25.45
49.15
10.74
3.53
15.37
70.36
7.26
3.47
15.53
73.74
7.22
3.46
15.52
73.8
45
45.6
45.4
50.2
50.4
50.7
44.8
59.5
65.6
66.2
66.3
66.8
14.3
24.9
26
28.1
28.3
28.4
13.14
15.74
29.22
41.9
45.96
1.77
12.21
40.06
15.3
13.86
28.94
41.9
12.39
8.57
23.12
55.92
11.65
8.52
23.2
56.63
11.04
8.25
23.04
57.68
11.5
49.4
63.4
67.2
67.2
67.2
25.7
63.5
77.6
77
76.9
76.9
4.8
28.3
42.6
43.5
43.5
43.5
6.2
11.62
33.3
48.88
29.67
2.89
18.13
49.31
11.62
2.91
13.38
72.09
8.31
2.27
12.01
77.41
8.28
2.26
11.96
77.5
8.3
2.25
11.95
77.5
0.75<AED<=1
0.5<AED<=0.75
0.25<AED<=0.5
0<=AED<=0.25


(A)
(B)
(C)
Uni-exon transcripts
Transcripts without UTRs
Transcripts with UTRs
Transcripts with micro-exons
Transcripts with long introns
Transcripts on same strand within 250 nucleotides of each other
Transcripts on opposite strand within 250 nucleotides of each other
Genes with more than 1 transcript
Genes with 1 transcript
No Data Available
#
0
10
20
30
40
50
Transcript F1 score
0
10
20
30
Transcript F1 score
10
20
30
40
50
Transcript F1 score
BRAKER2
MAKER2
PASA
FINDER
FINDER_BRAKER2
FINDER_BRAKER2_PROTEIN


Percentage of transcripts correctly detected
by each annotation pipeline
80
60
40
20
0
#
#
#
#
#
5 star
(9067)
4 star
(18374)
3 star
(2847)
2 star
(2665)
1 star
(1823)
🟥 BRAKER2
🟦 MAKER2
🟧 PASA
🟩 FINDER



Logarithm of distance of Predicted TSS
from annotated TSS
0
1
2
3
4
***
ns
***
Arabidopsis thaliana
Oryza sativa
Zea mays
🟥 BRAKER2
🟦 FINDER


(A)
Arabidopsis thaliana (987)
Brassicaceae (889)
Mesangiospermae (935)
Magnoliopsida (1257)
Embryophyta (2510)
Streptophytina (716)
Streptophyta (1874)
Viridiplantae (2453)
Eukaryota (9589)
cellular organisms (26021)
0 10 20 30 40
Transcript F1 score
(B)
Oryza sativa (3074)
Oryza (1838)
Poaceae (1659)
commelinids (650)
Mesangiospermae (1060)
Magnoliopsida (1269)
Embryophyta (2015)
Streptophytina (581)
Streptophyta (1547)
Viridiplantae (1922)
Eukaryota (6453)
cellular organisms (18913)
0 10 20 30 40
Transcript F1 score
(C)
Zea mays (2421)
Poaceae (682)
Mesangiospermae (948)
Magnoliopsida (1586)
Embryophyta (2444)
Streptophytina (905)
Streptophyta (2096)
Viridiplantae (2778)
Eukaryota (10092)
cellular organisms (23997)
0 10 20 30 40 50
Transcript F1 score
#
🟥 BRAKER
🟦 MAKER
🟧 PASA
🟩 FINDER
🟪 FINDER1_BRAKER
▨ FINDER1_BRAKER_PROTEIN

	Arabidopsis thaliana							*Oryza sativa*							*Zea mays (NCBI)*					
	BRAKER2	MAKER2	PASA	FINDER	FINDER+BRAKER2	FINDER+BRAKER2+PROTEIN		BRAKER2	MAKER2	PASA	FINDER	FINDER+BRAKER2	FINDER+BRAKER2+PROTEIN		BRAKER2	MAKER2	PASA	FINDER	FINDER+BRAKER2	FINDER+BRAKER2+PROTEIN
Base Specificity	**91.08**	74.87	62.71	74.46	75.01	75.04		**57.63**	52.67	36.77	42.46	42.6	42.75		6.52	45.22	60.39	**62.97**	**62.96**	**62.92**
Base Sensitivity	60.27	52.55	69.41	**71.45**	**74.14**	**74.23**		36.85	40.17	59.19	**61.43**	**61.83**	**62.37**		50.43	54.38	66.7	**72.1**	**72.15**	**72.16**
Base F1 score	72.54	61.76	65.89	**72.92**	**74.57**	**74.63**		44.95	45.58	45.36	**50.21**	**50.44**	**50.73**		11.55	49.38	63.39	**67.23**	**67.24**	**67.22**
Exon Specificity	80.28	**95.35**	90.74	91.79	91.43	91.43		38.4	**72.31**	67.37	67.67	67.5	67.62		15.96	65.14	**84.26**	79.71	79.59	79.57
Exon Sensitivity	73.29	55.04	67.49	**69.93**	**71.62**	**71.7**		53.74	50.59	63.99	**64.81**	**65.22**	**66.06**		65.65	61.84	71.84	**74.38**	**74.4**	**74.4**
Exon F1 score	76.63	69.79	77.41	**79.38**	**80.32**	**80.37**		44.79	59.53	65.64	**66.21**	**66.34**	**66.83**		25.68	63.45	**77.56**	76.95	76.91	76.9
Intron Specificity	86	**98.26**	95.98	96.55	96.26	96.26		52.8	**76.87**	72.36	73.24	73.19	73.26		21.77	74.41	**88.45**	84.86	84.78	84.76
Intron Sensitivity	84.25	60.64	76.11	**76.51**	**78.33**	**78.42**		**76.78**	55.77	70.44	71.03	71.52	72.34		80.88	64.99	75.19	**78.12**	**78.15**	**78.15**
Intron F1 score	85.12	75	84.9	**85.37**	**86.37**	**86.43**		62.57	64.64	71.39	**72.12**	**72.35**	**72.8**		34.31	69.38	**81.28**	**81.35**	**81.33**	**81.32**
Transcript Specificity	49.91	**76.9**	55.21	60.04	59.82	59.82		12.71	**40.13**	23.17	24.54	24.59	24.54		2.75	32.67	**48.6**	44.78	44.84	44.78
Transcript Sensitivity	30.26	21.74	28.62	**37.21**	**39.21**	**39.28**		16.35	18.01	29.51	**33**	**33.32**	**33.82**		19.57	25.01	37.94	**42.25**	**42.3**	**42.31**
Transcript F1 score	37.68	33.9	37.7	**45.95**	**47.37**	**47.42**		14.3	24.86	25.96	**28.15**	**28.3**	**28.44**		4.82	28.33	42.61	**43.48**	**43.53**	**43.51**
Gene Specificity	51.58	**76.9**	61.79	67.33	66.56	66.55		13.46	**40.13**	36.81	32.79	32.76	32.38		2.78	32.67	**55.69**	52.3	52.29	52.19
Gene Sensitivity	50.43	37.2	44.39	57.96	**61.3**	**61.4**		18.07	19.4	31.37	**35.05**	**35.4**	**35.96**		29.02	38.14	51.57	**57.35**	**57.42**	**57.43**
Gene F1 score	51	50.14	51.66	**62.29**	**63.82**	**63.87**		15.43	26.16	33.87	**33.88**	**34.03**	**34.08**		5.07	35.19	53.55	**54.71**	**54.74**	**54.68**

Organism	Number of transcript models borrowed from BRAKER	Percentage of transcript models borrowed from BRAKER	Improvement in average annotation score	Number of transcript models from protein alignments	Percentage of transcript models from protein alignments	Improvement in average annotation score
Arabidopsis thaliana	1692	5	1.43	185	0.01	0.05
Oryza sativa	5662	10	0.15	440	0.01	0.15
Zea mays	1061	2	0.05	452	0.01	-0.02
Caenorhabditis elegans	4807	18	0.48	389	0.01	0.58
Drosophila melanogaster	2421	9	0.44	481	0.02	0.22
Homo sapiens	5776	16	0.05	229	0.01	0.15
Hordeum vulgare	1065	3	0.01	19	0	-0.57

	Name	Description
Group 1	Uni-exon transcripts	Transcripts having a single exon and no introns
Group 2	Transcripts without UTRs	Transcripts missing either the 5' or the 3' UTR sequence
Group 3	Transcripts with UTRs	Transcripts having both UTRs
Group 4	Transcripts with micro-exons	Transcripts where at least one exon has length less than 50 nucleotides
Group 5	Transcripts with long introns	Transcripts where at least one intron has a length greater than 10,000 bp
Group 6	Closely placed transcripts on same strand	Transcripts on the same strand having less than 250 nucleotides between each other
Group 7	Closely placed transcripts on opposite strand	Transcripts on the opposite strands having less than 250 nucleotides between each other
Group 8	Multi transcript gene	Transcripts of a gene that have multiple transcripts
Group 9	Single transcript gene	Transcripts of a gene that have single transcript

	Arabidopsis thaliana						*Oryza sativa*						*Zea mays (RefSeq)*				
	STRAWBERRY	SCALLOP	STRINGTIE	PSICLASS	FINDER		STRAWBERRY	SCALLOP	STRINGTIE	PSICLASS	FINDER		STRAWBERRY	SCALLOP	STRINGTIE	PSICLASS	FINDER
Base Specificity	38.41	37.34	58.35	62.63	**74.46**		22.6	24.23	39.83	36.64	**42.46**		30.06	29.6	49.33	55.61	**62.97**
Base Sensitivity	**87.06**	85.3	80.22	70.83	71.45		**78.2**	77.64	70.87	60.27	61.43		**81.2**	79.08	76.98	70.52	72.1
Base F1 score	53.3	51.94	67.56	66.48	**72.92**		35.07	36.93	51	45.57	**50.21**		43.88	43.08	60.13	62.18	**67.23**
Exon Specificity	43.86	70.64	74.82	89.82	**91.79**		23.51	42.97	51.7	66.29	**67.67**		37.18	52.37	60.33	77.76	**79.71**
Exon Sensitivity	**85.3**	79.67	79.29	69.54	69.93		79.08	76.65	75.47	**65.75**	64.81		**85.03**	81.68	81.77	75.88	74.38
Exon F1 score	57.93	74.88	76.99	78.39	**79.38**		36.24	55.07	61.36	66.02	**66.21**		51.74	63.82	69.43	76.81	**76.95**
Intron Specificity	55.32	78.7	80.58	95.29	**96.55**		29.13	48.56	56.5	71.41	**73.24**		43.79	56.74	64.79	82.8	**84.86**
Intron Sensitivity	**92.06**	89.99	87.75	77.63	76.51		85.84	**85.41**	83.28	71.72	71.03		**90.19**	86.69	86.05	78.99	78.12
Intron F1 score	69.11	83.97	84.01	**85.56**	85.37		43.5	61.92	67.32	71.56	**72.12**		58.96	68.59	73.92	80.85	**81.35**
Transcript Specificity	6.88	24.84	35.02	56.82	**60.04**		1.59	9.03	14.26	24.22	**24.54**		6.96	17.8	26.23	43.96	**44.78**
Transcript Sensitivity	31.68	32.19	35.23	31.8	**37.21**		26.69	29.43	31.2	32.37	**33**		**48.71**	46.71	47.76	42.48	42.25
Transcript F1 score	11.3	28.04	35.12	40.78	**45.95**		3	13.82	19.57	27.71	**28.15**		12.18	25.78	33.86	43.21	**43.48**
Gene Specificity	33.16	36.51	63.7	65.67	**67.33**		17.61	20.16	35.69	**40.39**	32.79		21.83	26.33	47.43	**54.04**	52.3
Gene Sensitivity	44.77	46.18	50.59	49.33	**57.96**		28.91	31.79	33.65	34.54	**35.05**		55.2	55.85	58.03	57.22	**57.35**
Gene F1 score	38.1	40.78	56.39	56.34	**62.29**		21.89	24.67	34.64	**37.24**	33.88		31.29	35.79	52.2	**55.58**	54.71

	Number of FINDER1 transcripts having TSS better than BRAKER2	Number of BRAKER2 transcripts having TSS better than FINDER1
Arabidopsis thaliana	**15063 (65%)**	8022 (35%)
Oryza sativa	**11089 (66%)**	5762 (34%)
Zea mays (NCBI)	**20721 (76%)**	6628 (24%)
Zea mays (AGPv3)	7618 (28%)	**19731 (72%)**
Zea mays (AGPv4)	**18114 (69%)**	8297 (31%)
Caenorhabditis elegans	8681 (33%)	**17730 (67%)**
Drosophila melanogaster	**10238 (63%)**	5917 (37%)
Homo sapiens	**10158 (74%)**	3486 (26%)
Hordeum vulgare	**10373 (65%)**	5607 (35%)