Enhancing eukaryotic gene structure annotation via changepoint analysis of short-read coverage data

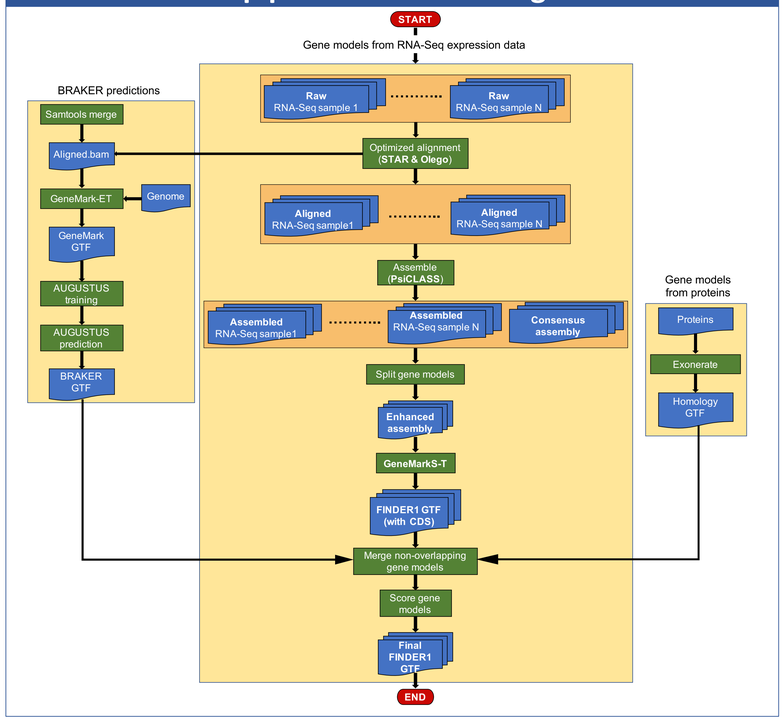
Sagnik Banerjee[3,4], Priyanka Bhandary[3,5], Margaret Woodhouse[1], Taner Z. Sen[2], Roger P. Wise[1,3,6,7], Carson Andorf*[1,3,8]

[1]USDA-ARS Corn Insects and Crop Genetics Research Unit, [2]USDA-ARS Crop Improvement and Genetics Research Unit, [3]Interdepartmental Bioinformatics and Computational Biology Program, Iowa State University, [4]Department of Statistics, Iowa State University, [5]Genetics, Developmental and Cell Biology, Iowa State University, [6]Interdepartmental Genetics & Genomics program, Iowa State University, [7]Department of Plant Pathology and Microbiology, Iowa State University, [8]Department of Computer Science, Iowa State University

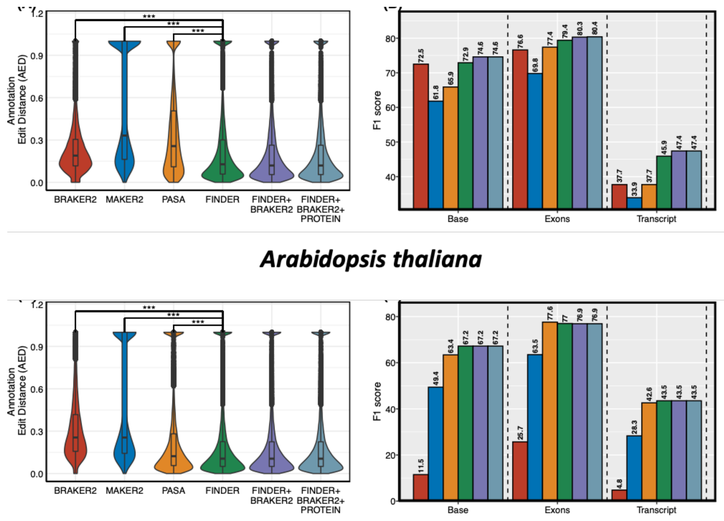
Abstract

Annotating eukaryotic genes is a non-trivial task that requires meticulous analysis of expression data. The presence of alternative splicing and overlapping genes increases this complexity. Currently available software annotates genomes from full-length cDNA or predict gene models from a database of splice junctions. Such approaches are highly sensitive to the quality of transcripts. Also, gene predictors (like BRAKER2) typically annotate only the coding regions of gene and report very few transcripts of a gene. To overcome these challenges, we have designed FINDER, which automates expression data download , read alignment, transcript assembly and gene prediction. FINDER is optimized to conduct read mapping with different settings to capture all biologically relevant alignments with special attention to micro-exons (exon length of 50 nucleotides or fewer). It integrates prediction results from BRAKER2 with assemblies constructed from expression data to approach the goal of exhaustive genome annotation. We tested FINDER on *Arabidopsis thaliana* and *Zea mays*. In each case, FINDER reported an improved transcript sensitivity and specificity compared with BRAKER2, MAKER2 and PASA. It also reports transcripts and recognizes genes expressed under specific conditions, flags problematic annotations, provides alternative annotations when conflicting evidence exists and identifies the best representative transcriptome.

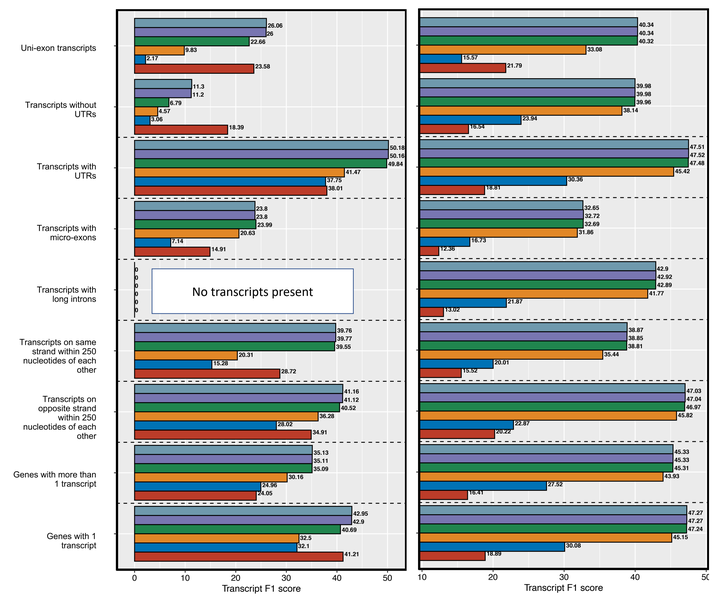
Informatics pipeline to construct gene models



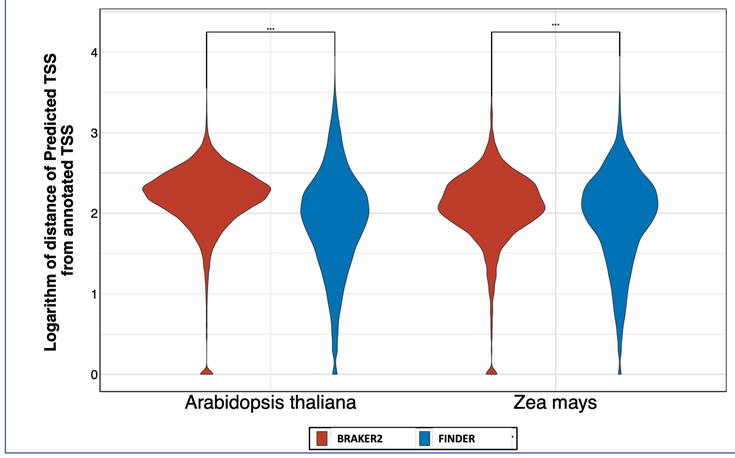
FINDER generates significantly better gene structure annotations



FINDER produces better gene models on different gene groups



Transcription start site (TSS) of FINDER gene models are closer to reference TSS as compared to BRAKER2



Changepoint detection enables correct annotation of gene models using read coverage

Conclusion

- FINDER detects UTR sequences in addition to CDS sequences
- FINDER detects more transcripts per gene compared to other annotation pipelines.
- Changepoint detection improves gene assemblies
- Combining predicted gene models with those constructed from RNA-Seq data boosts transcript discovery

Acknowledgements

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 disbursed through United Stated Department of Agriculture (USDA)

Contact

Sagnik Banerjee
Iowa State University
Email: sagnik@iastate.edu

References

1. Del Angel, V. D., et al. (2018). Ten steps to get started in Genome Assembly and Annotation. *F1000Research* 7.
2. Hoff, K. J., et al. (2016). BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS: Table 1. *Bioinformatics* 32, 767–769. doi:10.1093/bioinformatics/btv661.
3. Holt, C., et al. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12, 491. doi:10.1186/1471-2105-12-491
4. Kawahara, Y., et al. (2009). Change-point detection in time-series data by direct density-ratio estimation. in *Proceedings of the 2009 SIAM International Conference on Data Mining* (SIAM), 389–400.
5. Kawahara, Y., et al. (2007). Change-point detection in time-series data based on subspace identification. in *Seventh IEEE International Conference on Data Mining (ICDM 2007)* (IEEE), 559–564.
6. Liu, R., et al.(2017). Strawberry: Fast and accurate genome-guided transcript reconstruction and quantification from RNA-Seq. *PLoS Comput. Biol.* 13, e1005851.
7. Pertea, M., et al. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* 33, 290–295. doi:10.1038/nbt.3122
8. Shao, M., et al. (2017). Accurate assembly of transcripts through phase-preserving graph decomposition. *Nat. Biotechnol.* 35, 1167–1169. doi:10.1038/nbt.4020.
9. Song, L., et al. (2019). A multi-sample approach increases the accuracy of transcript assembly. *Nat. Commun.* 10, 5000. doi:10.1038/s41467-019-12990-0.