BRAKER2: A pipeline integrating data on genomic, RNA and protein sequences into inference of plant and animal genome annotation

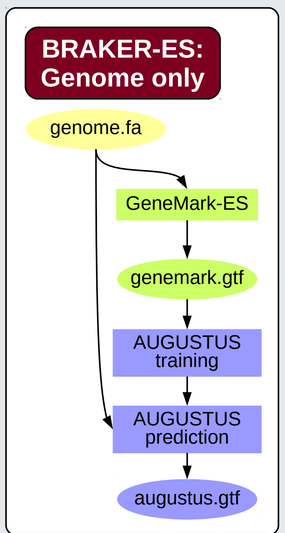
Katharina J. Hoff[1*], Tomáš Brůna[2*], Alexandre Lomsadze[2*], Mario Stanke[1] and Mark Borodovsky[2]

1) Institute for Mathematics and Computer Science, University of Greifswald, Greifswald, GERMANY
2) Joint Georgia Tech and Emory Wallace H Coulter Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, U.S.A.
Contact: katharina.hoff@uni-greifswald.de, *) Authors contributed equally

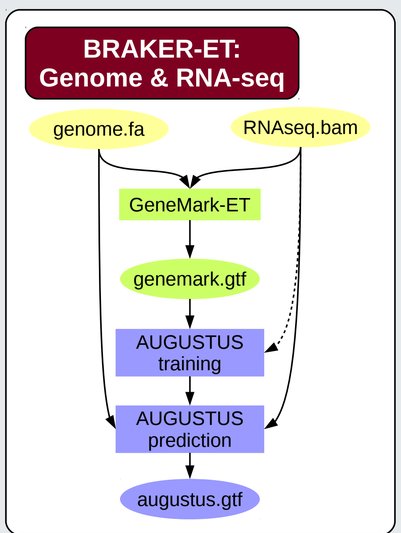
Abstract

The rapidly growing number of sequenced eukaryotic genomes requires fully automated methods for accurate gene structure annotation. With this goal in mind, we had developed BRAKER1 [1], a combination of self-training GeneMark-ET [2] and AUGUSTUS [3], that uses genomic and RNA-seq data to automatically generate full gene structure annotations in novel genomes (including alternative isoforms).
BRAKER2 [4] is an extension of BRAKER1. The new tool supports diverse annotation modes: i/ use of genome sequence only (ES), ii/ use of genome and RNA-seq data (ET), iii/ use of genome and proteins of possibly distant evolutionary origin (EP), iv/ use of genome, RNA-seq data and proteins. We have assessed gene prediction accuracy of BRAKER2 for two model organisms *Caenorhabditis elegans* and *Drosophila melanogaster*. In addition, we applied BRAKER2 to ten non-model organisms, using VARUS [5] as RNA-seq sampling tool and OrthoDB [6] as protein data resource.
BRAKER2 is available for download at http://github.com/Gaius-Augustus/BRAKER, the GeneMark-EP protein mapping pipeline is available for download at http://exon.gatech.edu/GeneMark/Braker/protein_mapping_pipeline.tar.gz.

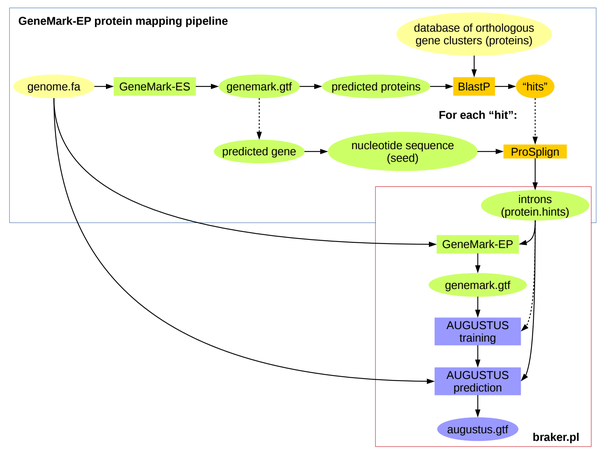
BRAKER-ES: Genome only



BRAKER-ET: Genome & RNA-seq



BRAKER-EP: Genome & proteins



BRAKER-ETP: Genome, RNA-seq & proteins

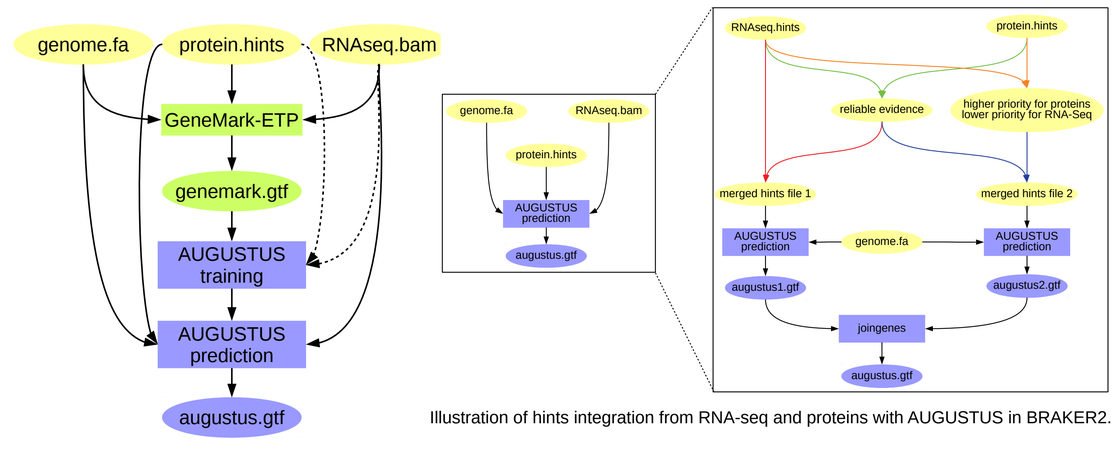


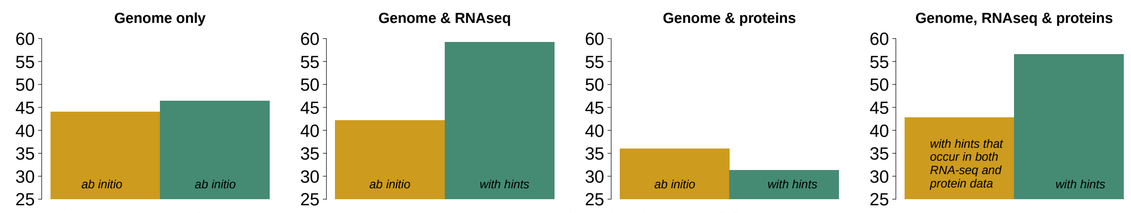
Illustration of hints integration from RNA-seq and proteins with AUGUSTUS in BRAKER2.

Results for model species

$$\text{Gene Level Accuracy } F1 = \frac{2 * Sensitivity * Specificity}{Sensitivity + Specificity}$$

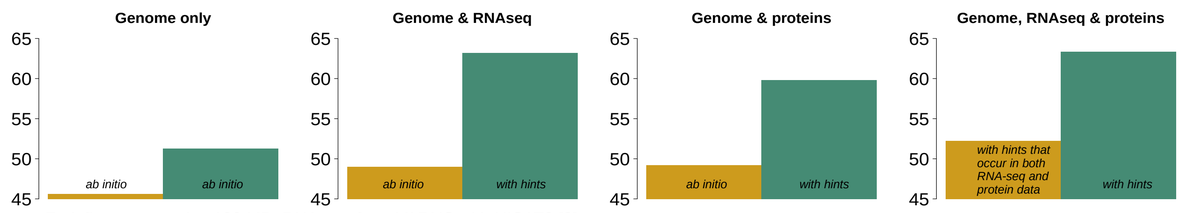
GeneMark
AUGUSTUS

Caenorhabditis elegans



- Roundworm, genome size 98 MB, RNA-seq from RGASP assessment aligned with HISAT2 [8]
- OrthoDB 10 Metazoa: 3 species from the same order, 4 species from same the phylum, 32 species from the same kingdom, none of same genus

Drosophila melanogaster

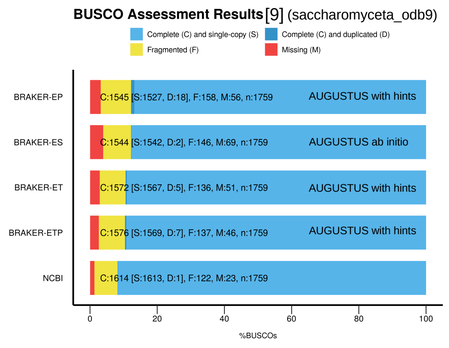


- Fruit fly, genome size 130 MB, RNA-seq from VARUS with HISAT2 [8]
- OrthoDB 10 Arthropoda: 36 species from the same order, 40 species from the same class, none of same genus

Results for non-model species

Verticillium dahliae

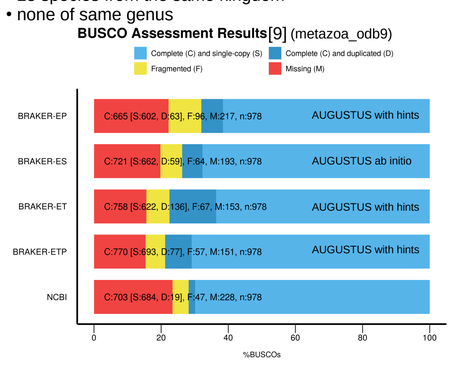
Genome: GCF_000150675.1_ASM15067v2_genomic.fna
Genome size: 33MB
RNA-seq: 50 M reads by VARUS with STAR [6]
Protein mapping pipeline with OrthoDB v10 Fungi:
- 12 species from the same order
- 75 species from the same class
- none of same genus



	BRAKER-ETP only	BRAKER-ETP & NCBI	NCBI only
Genes	4376	4921	5614
Exons	8267	18714	10771

Hymenolepsis microstoma

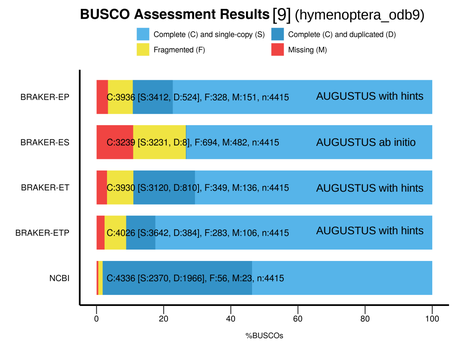
Genome: GCA_000469805.2_HMIC002_genomic.fna
Genome size: 176 MB
RNA-seq: 50 M reads by VARUS with HISAT2 [8]
Protein mapping pipeline with OrthoDB v10 Metazoa:
- 1 species from the same order
- 4 species from the same phylum
- 28 species from the same kingdom
- none of same genus



	BRAKER-ETP only	BRAKER-ETP & NCBI	NCBI only
Genes	7779	3647	5762
Exons	25860	48761	11781

Bombus terrestris

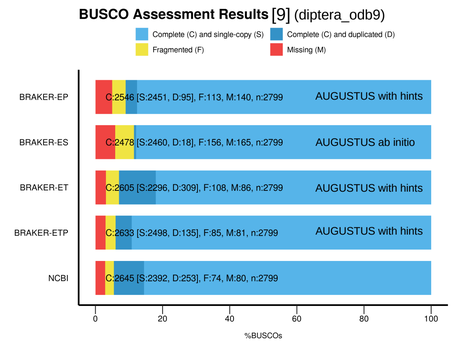
Genome: GCF_000214255.1_Bter_1.0_genomic.fna
Genome size: 241 MB
RNA-seq: 50 M reads by VARUS with HISAT2 [8]
Protein mapping pipeline with OrthoDB v10 Arthropoda:
- 6 species from the same family
- 39 species from the same order
- none of same genus



	BRAKER-ETP only	BRAKER-ETP & NCBI	NCBI only
Genes	10751	5041	5541
Exons	21159	62156	16186

Anopheles gambiae

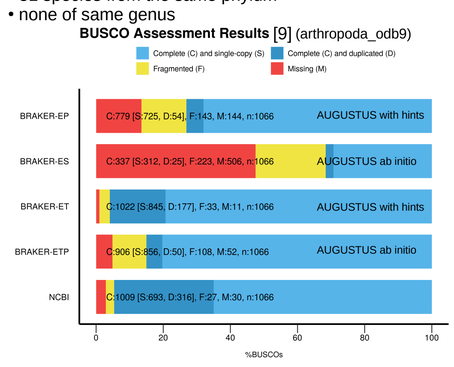
Genome: GCF_000005575.2_AgamP3_genomic.fna
Genome size: 257 MB
RNA-seq: 50 M reads by VARUS with HISAT2 [8]
Protein mapping pipeline with OrthoDB v10 Arthropoda:
- 3 species from the same family
- 42 species from the same order
- none of same genus



	BRAKER-ETP only	BRAKER-ETP & NCBI	NCBI only
Genes	9757	4624	7979
Exons	25578	33301	18508

Parasteatoda tepidariorum

Genome: GCF_000365465.2_Ptep_2.0_genomic.fna
Genome size: 1.4 GB
RNA-seq: 50 M reads by VARUS with HISAT2 [8]
Protein mapping pipeline with OrthoDB v10 Arthropoda:
- 1 species from the same order
- 10 species from the same class
- 31 species from the same phylum
- none of same genus

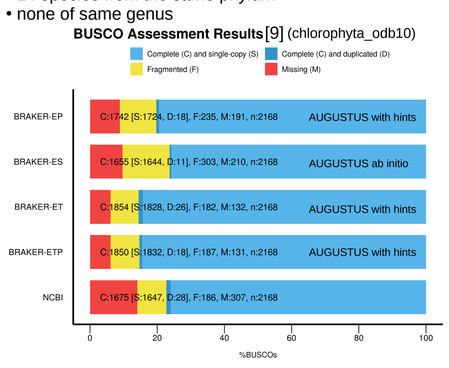


	BRAKER-ET* only	BRAKER-ET* & NCBI	NCBI only
Genes	34160	6812	11790
Exons	72681	117593	26212

*) BRAKER-ETP with hints was still running during poster printing, therefore we show results with BRAKER-ET instead.

Chlamydomonas rheinhardtii

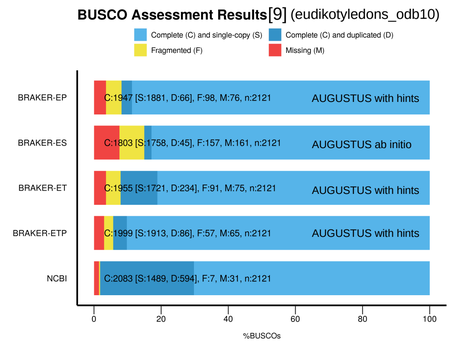
Genome: GCF_000002595.1_v3.0_genomic.fna
Genome size: 117 MB
RNA-seq: 50 M reads sampled by VARUS with STAR [6]
Protein mapping pipeline with OrthoDB v10 Plants:
- 3 species from the same order
- 4 species from the same class
- 14 species from the same phylum
- none of same genus



	BRAKER-ETP only	BRAKER-ETP & NCBI	NCBI only
Genes	13975	2242	12173
Exons	66449	69710	49584

Cucumis sativus

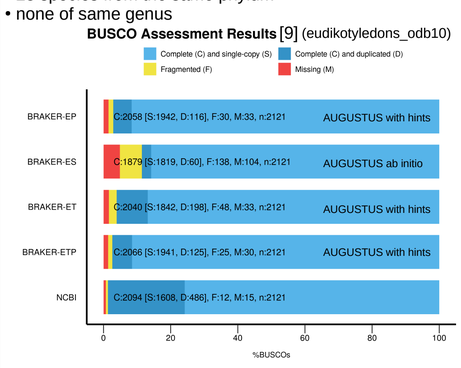
Genome: GCF_000004075.2_ASM407v2_genomic.fna
Genome size: 189 MB
RNA-seq: 50 M reads by VARUS with HISAT2 [8]
Protein mapping pipeline with OrthoDB v10 Plants:
- 2 species from the same family
- 23 species from the same phylum
- none of same genus



	BRAKER-ETP only	BRAKER-ETP & NCBI	NCBI only
Genes	16113	11269	7599
Exons	41986	99518	14300

Fragaria vesca

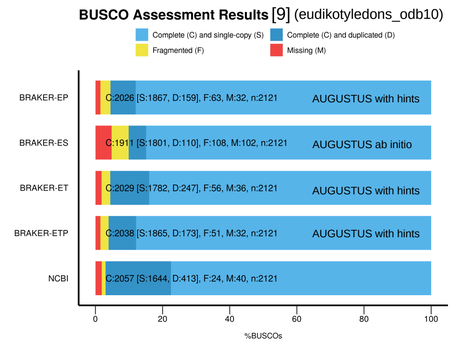
Genome: GCF_000184155.1_FraVesHawaii_1.0_genomic.fna
Genome size: 208 MB
RNA-seq: 50 M reads by VARUS with HISAT2 [8]
Protein mapping pipeline with OrthoDB v10 Plants:
- 4 species from the same family
- 8 species from the same order
- 23 species from the same phylum
- none of same genus



	BRAKER-ETP only	BRAKER-ETP & NCBI	NCBI only
Genes	22824	14432	9709
Exons	55920	110960	18952

Medicago truncatula

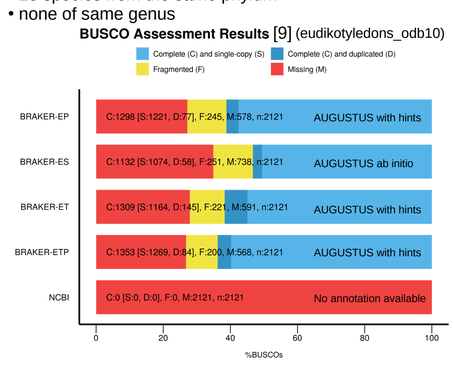
Genome: GCF_000219495.3_MedtrA17_4.0_genomic.fna
Genome size: 399 MB
RNA-seq: 50 M reads by VARUS with HISAT2 [8]
Protein mapping pipeline with OrthoDB v10 Plants:
- 9 species from the same family
- 12 species from the same phylum
- none of same genus



	BRAKER-ETP only	BRAKER-ETP & NCBI	NCBI only
Genes	36905	19201	31319
Exons	86670	144978	67874

Cannabis sativa

Genome: GCA_001865755.1_ASM186575v1_genomic.fna
Genome size: 566 MB
RNA-seq: 50 M reads by VARUS with HISAT2 [8]
Protein mapping pipeline with OrthoDB v10 Plants:
- 4 species from the same family
- 8 species from the same order
- 23 species from the same phylum
- none of same genus



	BRAKER-ETP only	BRAKER-ETP & NCBI	NCBI only
Genes	94908	NA	NA
Exons	294876	NA	NA

References

[1] Hoff, Katharina J., et al. "BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS." Bioinformatics 32.5 (2015): 767-769.
[2] Lomsadze, Alexandre, Paul D. Burns, and Mark Borodovsky. "Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm." Nucleic acids research 42.15 (2014): e119-e119.
[3] Stanke, Mario, et al. "Using native and syntenically mapped cDNA alignments to improve de novo gene finding." Bioinformatics 24.5 (2008): 637-644.
[4] Hoff, Katharina J., et al. "Whole-Genome Annotation with BRAKER" Springer Protocols (2019), in press.
[5] Stanke, Mario et al. "Automatic Genome Annotation Looping over Species" Poster PE0094 at PAG XXVII (2019).
[6] Kriventseva, Evgenia V., et al. "OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic acids research (2018): doi.org/10.1093/nar/gky1053s.
[7] Dobin, Alexander, et al. "STAR: ultrafast universal RNA-seq aligner." Bioinformatics 29.1 (2013): 15-21.
[8] Daehwan, Kim, et al. "HISAT: a fast spliced aligner with low memory requirements." Nature methods 12.4 (2015): 357.
[9] Waterhouse, Robert M., et al. "BUSCO applications from quality assessments to gene prediction and phylogenomics." Molecular biology and evolution 35.3(2017):543-548