BRAKER2: Incorporating Protein Homology Information into Gene Prediction with GeneMark-EP and AUGUSTUS

A pipeline for fully automated training and prediction

Plant and Animal Genomes XXVI, January 14th 2018

Katharina J. Hoff,
Alexandre Lomsadze,
Mario Stanke,
Mark Borodovsky

Presenting author: katharina.hoff@uni-greifswald.de

Contents

1. **Gene prediction**
2. **BRAKER1: RNAseq**
3. **BRAKER2: proteins**
   - Short evolutionary distance
   - Long evolutionary distance
4. **Summary**
5. **References**

Structural genome annotation problem

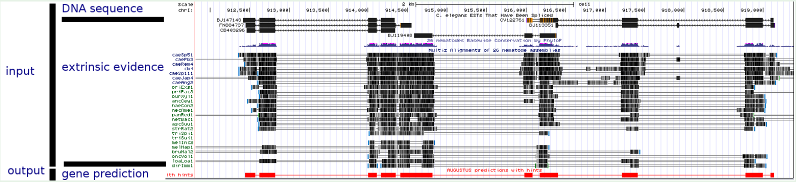
Input

- genome assembly
- extrinsic evidence, e.g. from RNAseq, protein database

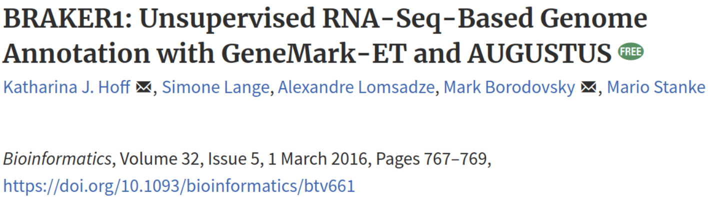
Output

- protein-coding genes: exon-intron structures (`.gff`)

Example (from Chr I in *C. elegans*)



BRAKER1: RNAseq integration



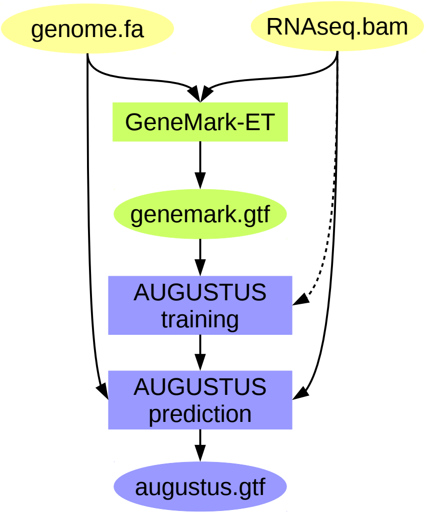
BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS FREE

Katharina J. Hoff ✉, Simone Lange, Alexandre Lomsadze, Mark Borodovsky ✉, Mario Stanke

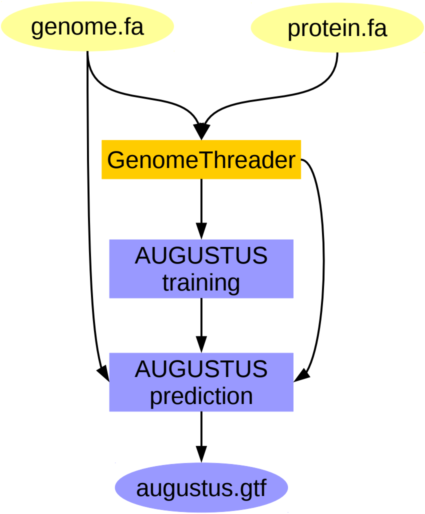
Bioinformatics, Volume 32, Issue 5, 1 March 2016, Pages 767–769, https://doi.org/10.1093/bioinformatics/btv661

- >4000 downloads
- 73 citations since 2016 (google scholar)

BRAKER1: RNAseq integration



BRAKER2: Part I - proteins of closely related species



Drosophila melanogaster and relatives

For a given species,

- the average number of mutations per genomic site was computed from alignments of ortholog gene sequences (including introns).
- the protein identity was computed as average of identity values of the best `exonerate` hit found for each protein of this species against the *D. melanogaster* genome.

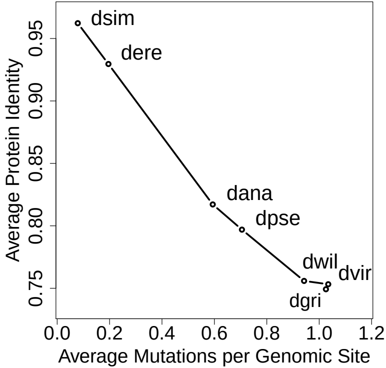
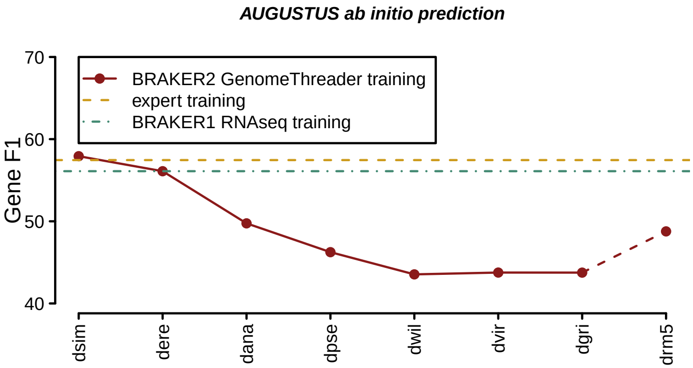
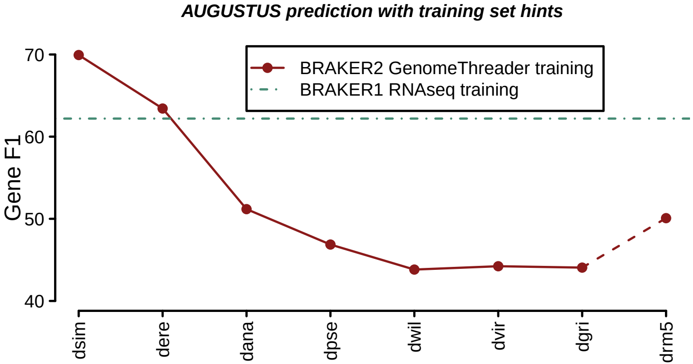


Image: S. König, L. Romoth, M. Stanke (2018) Comparative Genome Annotation

Increasing evolutionary distance leads to decreasing gene prediction accuracy of AUGUSTUS



Increasing evolutionary distance leads to decreasing gene prediction accuracy of AUGUSTUS



Increasing evolutionary distance leads to decreasing gene prediction accuracy of AUGUSTUS

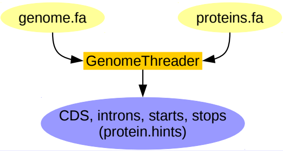
With increasing distance between query protein and target genome, spliced alignments become

- less sensitive while keeping a constant level of specificity (e.g. GenomeThreader),
- or both less sensitive and less specific (e.g. Exonerate).

Therefore, training AUGUSTUS on spliced alignments is suitable upon availability of a very closely related query species, only!

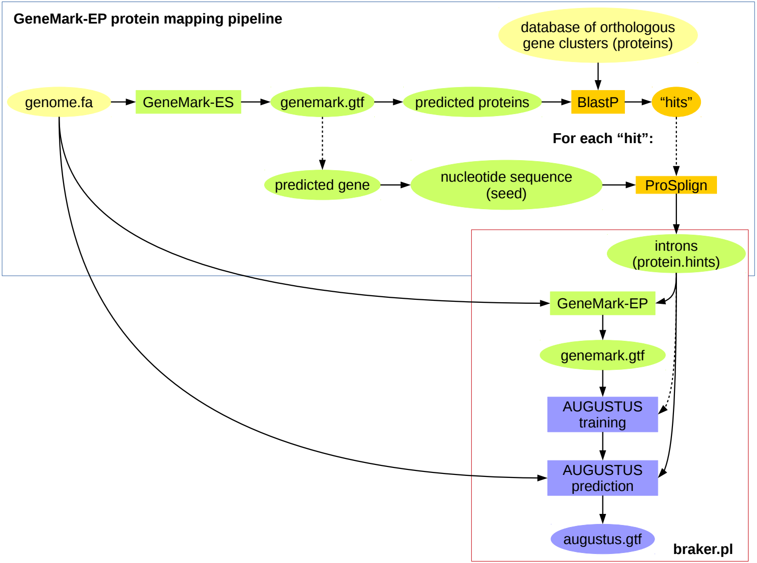
BRAKER2: Part II - proteins of more remote species

“Standard mapping approach”: proteins to genome



→ works well for closely related species, only

BRAKER2: Part II - proteins of more remote species



Protein database for gene prediction in *D. melanogaster*

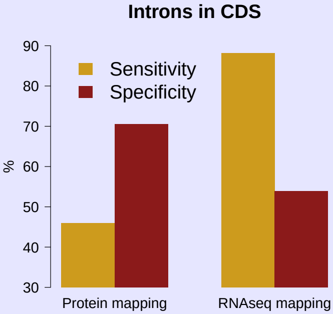
Insect portion of EggNOG (inNOG) excluding *Drosophila* species

- *Acyrthosiphon pisum*
- *Aedes aegypti*
- *Anopheles darlingi*
- *Anopheles gambiae*
- *Apis mellifera*
- *Atta cephalotes*
- *Bombyx mori*
- *Culex quinquefasciatus*
- *Danaus plexippus*
- *Heliconius melpomene*
- *Nasonia vitripennis*
- *Pediculus humanus*
- *Tribolium castaneum*

Intron recovery from protein mapping

Protein mapping with no *Drosophila* EggNOG (inNOG)

- 30,996 introns predicted
- 21,843 matched introns in CDS part of the annotated genes

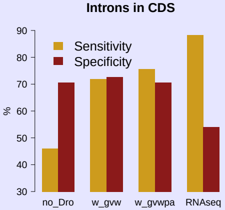


Mapping of proteins from remote species recovers ∼45% of introns with specificity of ∼70%.

Intron recovery from protein mapping

Protein mapping with some *Drosophila* species present as external evidence

no_Dro	no *Drosophila* species
w_gvw	with *D. grimshawi*, *D. virilis*, *D. willistoni*
w_gvwpa	with *D. grimshawi*, *D. virilis*, *D. willistoni*, *D. pseudoobscura*, *D. ananassae*

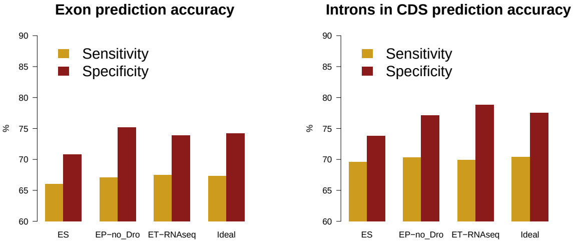


→ more introns were detected

→ performance of protein mapping with addition of 5 fly proteomes came closer to performance with RNAseq external evidence

Accuracy of GeneMark-EX with different sources of evidence

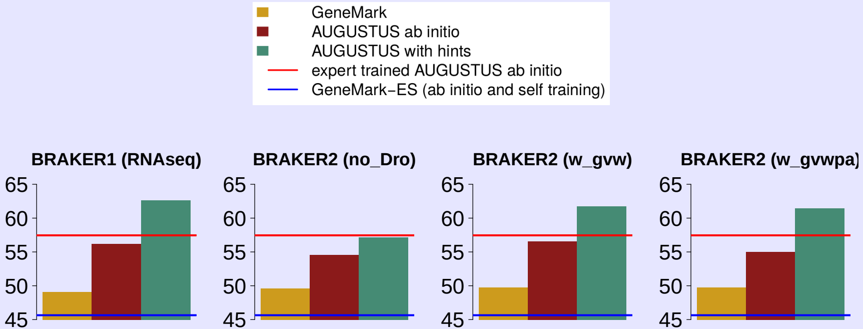
- results are on **softmasked genome** (strongly recommended!)



- GeneMark-EP and GeneMark-ET outperformed GeneMark-ES
- GeneMark-EP with “remote” proteins was comparable with GeneMark-ET
- GeneMark-EP and GeneMark-ET were close to the best possible performance: compared to training with “ideal” introns

Accuracy of BRAKER2

Gene prediction accuracy (F1)



Summary

- BRAKER2 is a novel fully automatic pipeline which makes gene prediction in eukaryotic genomes with RNAseq or protein external evidence.
- Training in BRAKER2 is done by GeneMark-EX which particularly can use remote proteins as external evidence.
- Prediction in BRAKER2 is done by AUGUSTUS using RNAseq or proteins as hints.

Ongoing & future work

- Optimization of evidence integration in BRAKER2
- Combining RNAseq and protein information
- UTR training & integration of RNAseq coverage information

References

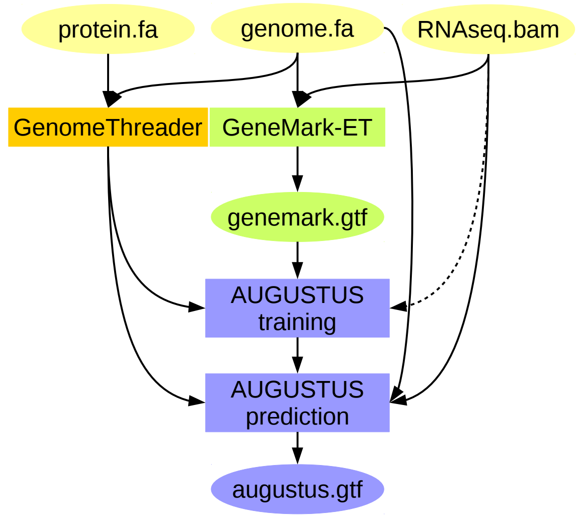
- Hoff, Katharina J., et al. “BRAKER1: unsupervised RNAseq-based genome annotation with GeneMark-ET and AUGUSTUS.” Bioinformatics 32.5 (2015): 767-769.
- Stanke, Mario, et al. “Using native and syntenically mapped cDNA alignments to improve de novo gene finding.” Bioinformatics 24.5 (2008): 637-644.
- Lomsadze, Alexandre, Paul D. Burns, and Mark Borodovsky. “Integration of mapped RNAseq reads into automatic training of eukaryotic gene finding algorithm.” Nucleic acids research 42.15 (2014): e119-e119.
- Slater, Guy St C., and Ewan Birney. “Automated generation of heuristics for biological sequence comparison.” BMC bioinformatics 6.1 (2005): 31.
- Gremme, Gordon. “GenomeThreader Gene Prediction Software.” (2014).
- Dobin, Alexander, et al. “STAR: ultrafast universal RNA-seq aligner.” Bioinformatics 29.1 (2013): 15-21.

BRAKER2 is available for download at

- `http://bioinf.uni-greifswald.de`
- `http://exon.gatech.edu`

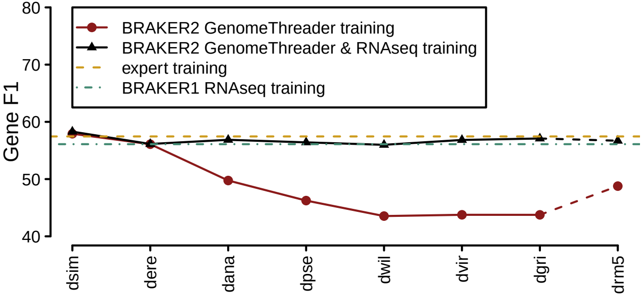
State of the art: BRAKER with RNAseq & proteins

Close homology

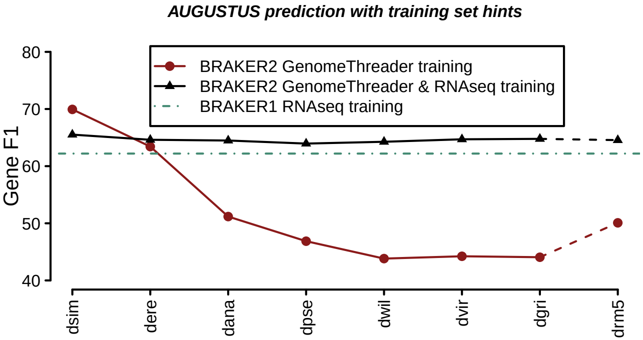


State of the art: BRAKER with RNAseq & proteins

AUGUSTUS ab initio prediction

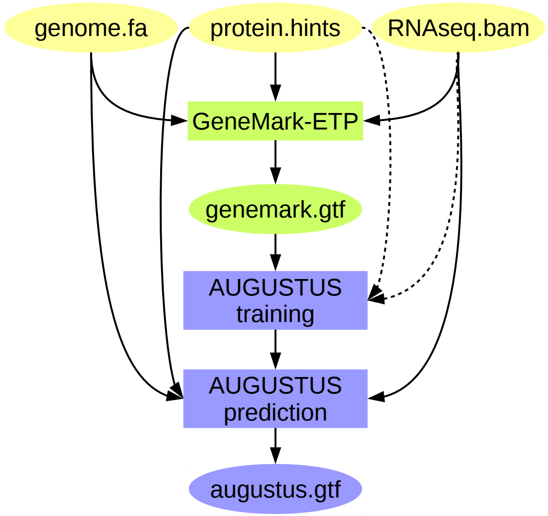


State of the art: BRAKER with RNAseq & proteins

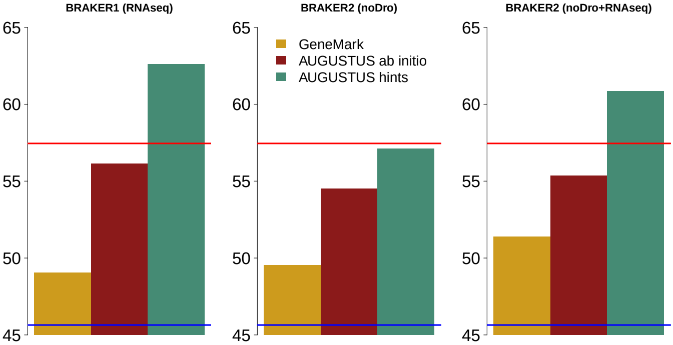


State of the art: BRAKER with RNAseq & proteins

Remote homology



State of the art: BRAKER with RNAseq & proteins


BRAKER1 (RNAseq)
BRAKER2 (noDro)
BRAKER2 (noDro+RNAseq)
GeneMark
AUGUSTUS ab initio
AUGUSTUS hints