Unsupervised RNA-Seq-based genome annotation with GeneMark-ET & AUGUSTUS

January 11th 2015

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

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

Contents

RNA-Seq & gene prediction

GeneMark-ET

AUGUSTUS

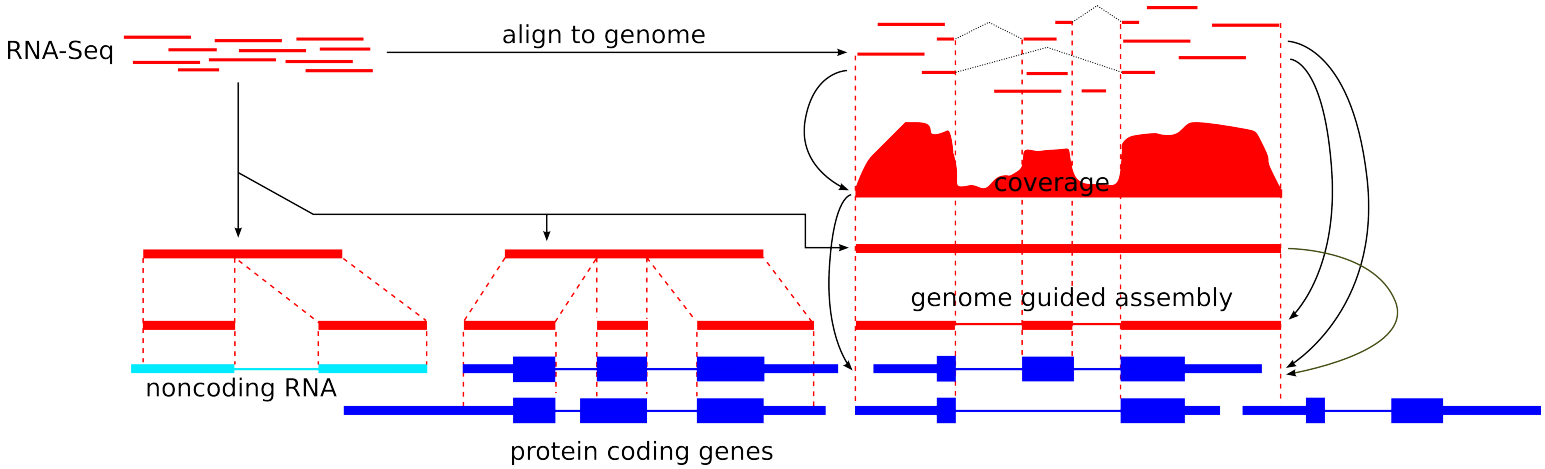
BRAKER1

Pipeline

Data sets

Results

From RNA-Seq to Genes



Approaches:

- de novo transcript assembly -> mapping -> gene prediction
- mapping -> genome-guided assembly -> gene prediction
- mapping -> gene prediction

GeneMark-ET uses RNA-Seq for Training

Nucleic Acids Research, 2014 **1**
doi: 10.1093/nar/gku557

Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm

Alexandre Lomsadze[1], Paul D. Burns[1] and Mark Borodovsky[1,2,3,*]

[1]Joint Georgia Tech and Emory Wallace H. Coulter Department of Biomedical Engineering, Atlanta, GA, USA 30332, [2]School of Computational Science and Engineering, Georgia Tech, Atlanta, GA, USA 30332 and [3]Department of Bioinformatics, Moscow Institute of Physics and Technology, Moscow, Russia 141700

- employs unsupervised training
- includes in training introns and exons anchored by mapped RNA-Seq reads
- does not require RNA-Seq reads assembly
- does not use RNA-Seq information in the *prediction* step

GeneMark-ET uses RNA-Seq for Training

Anchors from RNA-Seq for training

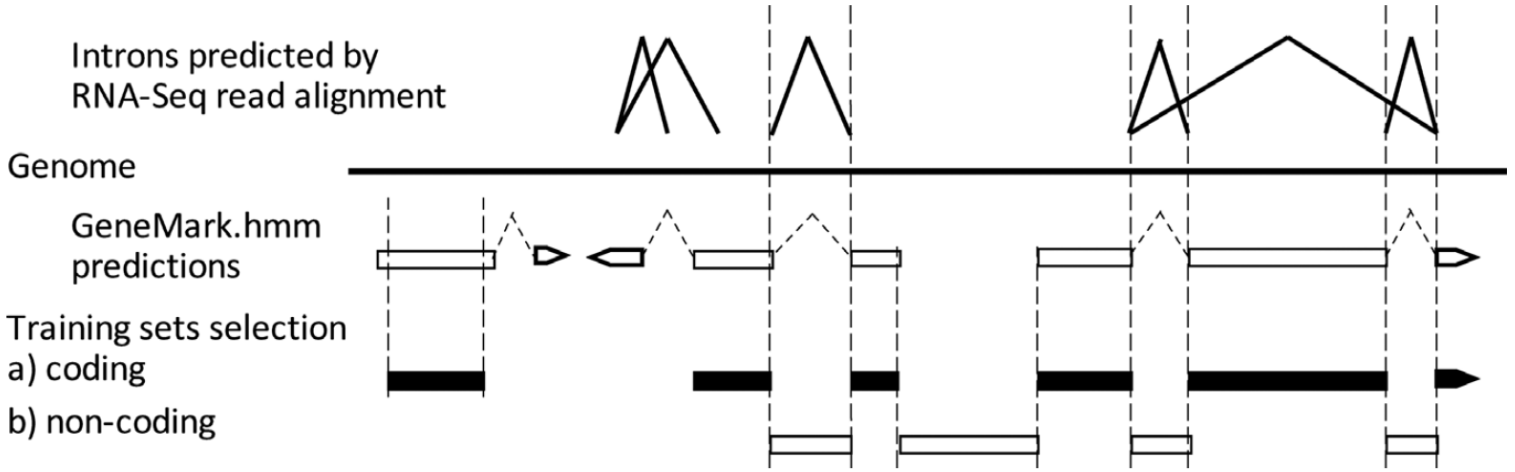
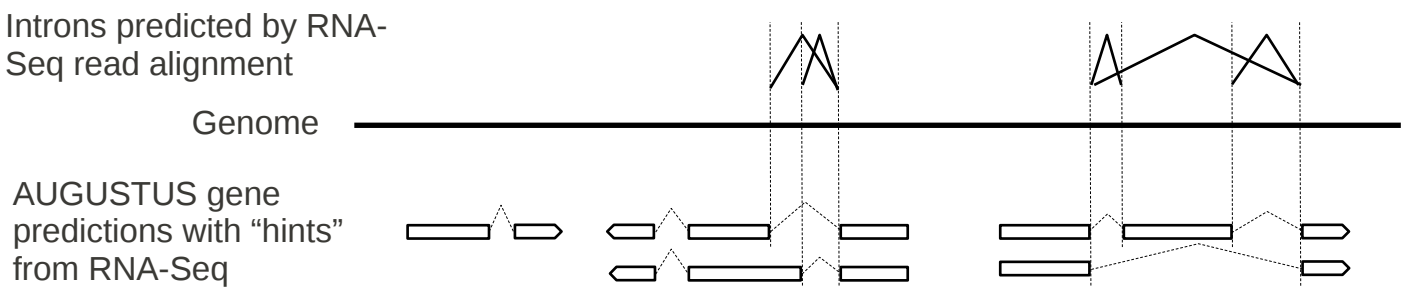


Figure 3. Selection of elements of training set in GeneMark-ET for the next iteration. The new training set of protein-coding regions is comprised from exons with at least one 'anchored splice site' as well as long exons predicted *ab initio* (>800 nt).

- employs unsupervised training
- includes in training introns and exons anchored by mapped RNA-Seq reads
- does not require RNA-Seq reads assembly
- does not use RNA-Seq information in the *prediction* step

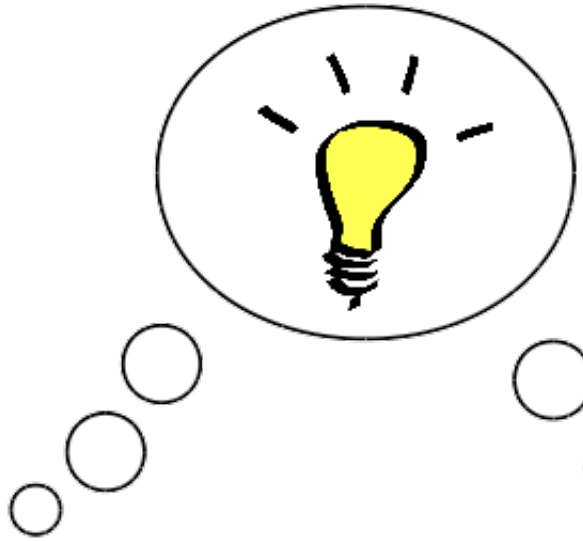
Simone Lange, Katharina J. Hoff, Alexandre Lomsadze, Mark Borodovsky, Mario Stanke

AUGUSTUS uses RNA-Seq for Prediction



- requires “prior data” for training
- uses intron information from RNA-seq for *prediction*
- no RNA-Seq assembly required

PAG 2014...



PAG 2014...

Our intention was to create a eukaryotic gene prediction tool that

- trains automatically
- improves state-of-the-art gene prediction accuracy
- uses RNA-Seq for training and prediction (as unassembled reads)
- is easy to use

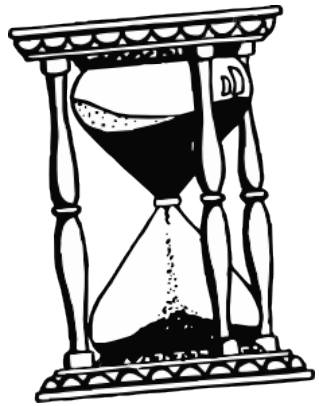
Simone Lange, Katharina J. Hoff, Alexandre Lomsadze, Mark Borodovsky, Mario Stanke

BRAKER1

BRAKER1

Running BRAKER1

```
braker.pl [OPTIONS] -genome=genome.fa -bam=rnaseq.bam
```



$\sim$ 1 day for fly on 1 CPU

Results

- BRAKER1-GeneMark-ET gene predictions
- BRAKER1-AUGUSTUS gene predictions

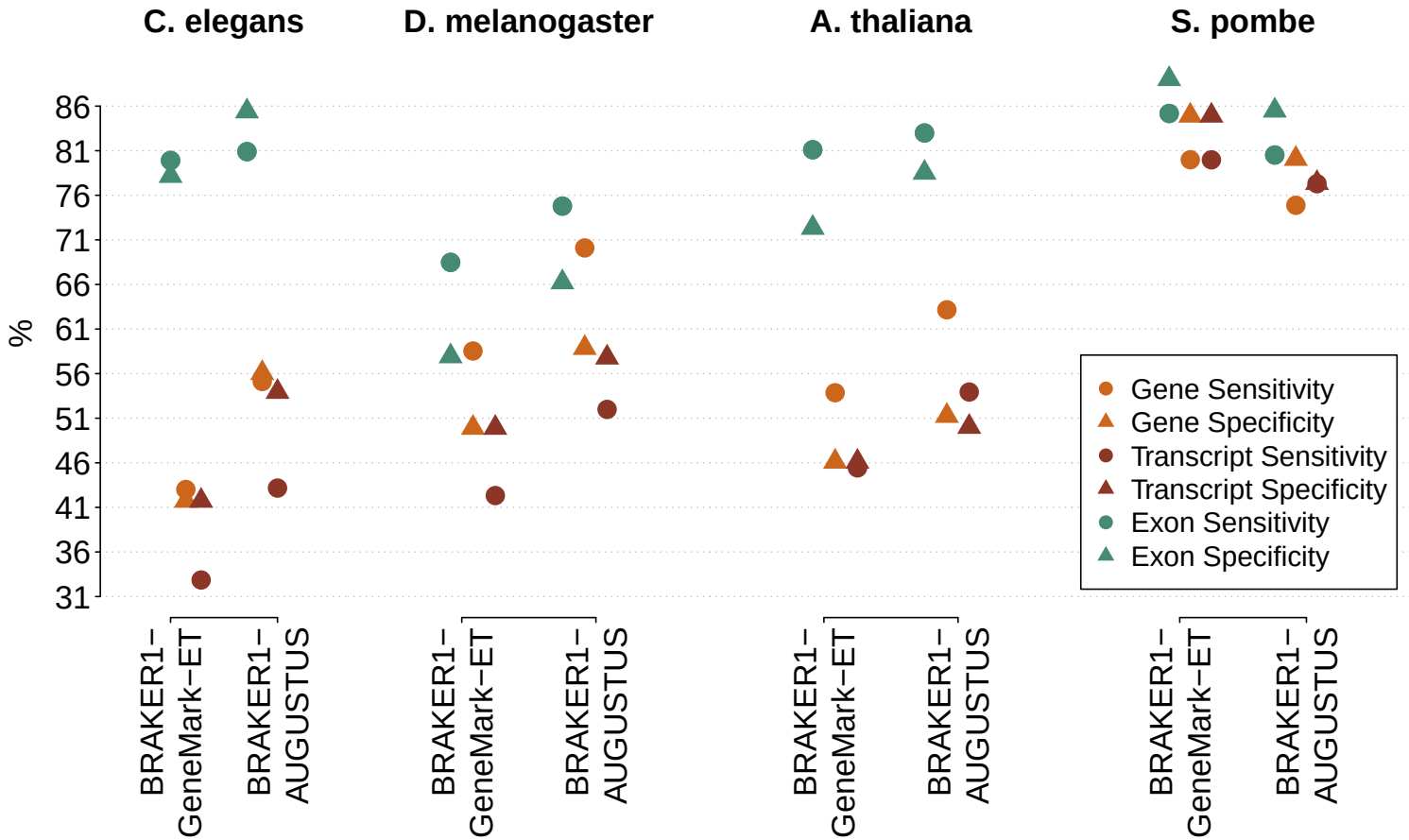
Data sets for accuracy evaluation

Model organisms and Illumina paired end libraries

- ***Drosophila melanogaster*** (flybase.org)
  - genome and reference annotation version R5
  - RGASP RNA-Seq libraries
- ***Arabidopsis thaliana*** (arabidopsis.org)
  - genome and reference annotation version TAIR 10
  - SRR934391
- ***Caenorhabditis elegans*** (wormbase.org)
  - genome and reference annotation version WS240
  - RGASP RNA-Seq library
- ***Schizosaccharomyces pombe*** (pombase.org)
  - genome and reference annotation version ASM294v2.23
  - SRR097898, SRR097899, SRR097900, SRR097902, SRR097903, SRR097905, SRR097906, SRR097907, SRR097908, SRR097909, SRR097912, SRR097915, SRR097917, SRR097921, SRR097922, SRR097925, SRR402833

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

Accuracy of BRAKER1



Comparing BRAKER1 to... MAKER2[1]?

Maker2

- generates “training genes” from assembled RNA-Seq
- uses GeneMark-ES, AUGUSTUS, SNAP
- integrates RNA-Seq evidence (assembled and reads) into gene prediction

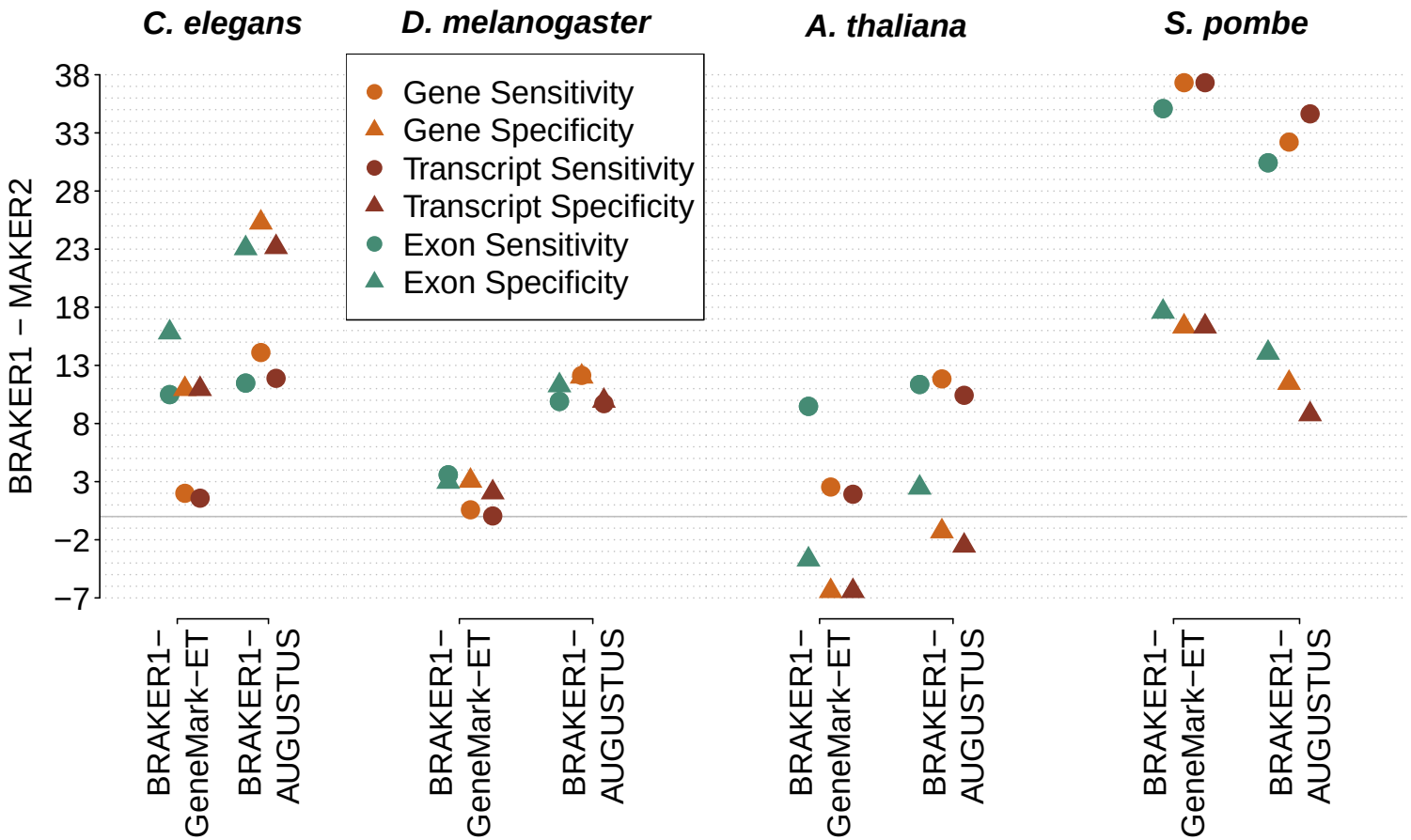
How we use MAKER2

- no protein database
- `keep_preds=1`
- include Cufflinks transcripts & read alignments
- MAKER2 masks repeats

[1] Following the tutorial at following tutorial at http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/MAKER_Tutorial_for_GMOD_Online_Training_2014

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

Comparing BRAKER1 to MAKER2



Future Work

- integration of protein information
- further optimization of BRAKER1 parameters
- UTR training & integration of RNA-Seq coverage information

BRAKER1 is available for download at

http://bioinf.uni-greifswald.de

and

http://exon.gatech.edu

Acknowledgements

Simone Lange,
Alexandre Lomsadze,
Mark Borodovsky,
Mario Stanke