Spring 2024

BIOS 477/877

Bioinformatics and Molecular Evolution

Lecture 30

1

TODAY'S TOPICS

➢ Gene prediction
- Ab initio & Combination methods

➢ Genome annotation pipelines

➢ Course Evaluation on Canvas
Submit by 11:59 PM (Saturday)

2

Gene prediction methods

➢ Various pattern recognition methods are used
- Decision trees [*e.g.*, MORGAN, GlimmerM]
- Discriminant function analysis (LDA, QDA)
  [*e.g.*, MZEF, FGENES, HEXON]
- Neural networks (NN) [*e.g.*, GRAIL II]
- Hidden Markov Models (HMMs)
  [*e.g.*, GENSCAN, HMMGene, GeneMark.hmm, Glimmer, Augustus]
- Combiner or ensemble methods
  [*e.g.*, JIGSAW, GLEAN, GeneComber, EvidenceModeler, TSEBRA]
- Combinations with similarity or evidence
  [*e.g.*, NSCAN, CONTRAST, Augustus, GeMoMa, Maker2, Braker3]

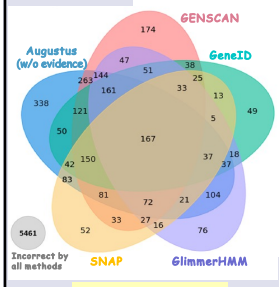
3

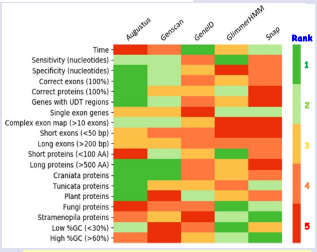
A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms

Nicolas Scalzitti, Anne Jeannin-Girardon, Pierre Collet, Olivier Poch & Julie D. Thompson

BMC Genomics **21**, Article number: 293 (2020) | Cite this article

[Benchmark data]
889 confirmed protein-coding genes, including 7,968 exons
Collected from diverse organisms (protists ~ human)





Correctly predicted exons

Sensitivity=TP/(TP+FN); Specificity=TP/(TP+FP)
UDT: undetermined (including 'N')

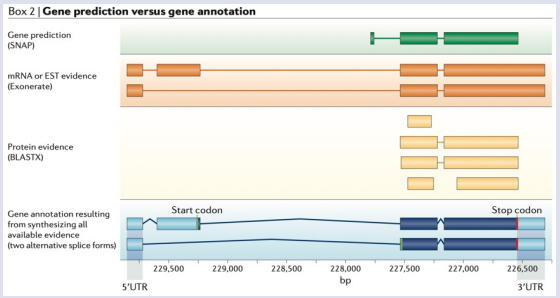
4

Gene prediction methods

➢ Various pattern recognition methods are used
- Decision trees [*e.g.*, MORGAN, GlimmerM]
- Discriminant function analysis (LDA, QDA)
  [*e.g.*, MZEF, FGENES, HEXON]
- Neural networks (NN) [*e.g.*, GRAIL II]
- Hidden Markov Models (HMMs)
  [*e.g.*, GENSCAN, HMMGene, GeneMark.hmm, Glimmer, Augustus]
- Combiner or ensemble methods
  [*e.g.*, JIGSAW, GLEAN, GeneComber, EvidenceModeler, TSEBRA]
- Combinations with similarity or evidence
  [*e.g.*, NSCAN, CONTRAST, Augustus, GeMoMa, Maker2, Braker3]

5

ab initio prediction + evidence

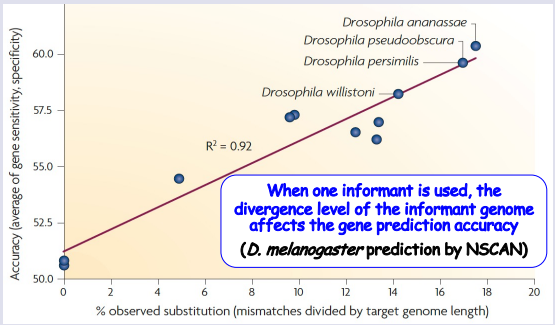


Yandell & Ence (2012)

Various evidence can be used to improve gene prediction results: *e.g.*, 5' and 3' untranslated regions (UTRs), alternatively spliced variants (isoforms)

6

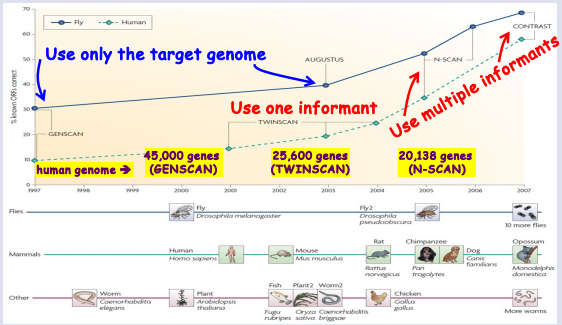
ab initio + similarity methods



Brent (2008)

7

ab initio + similarity methods



Brent (2008)

8

Simultaneous gene finding in multiple genomes

Stefanie König*, Lars W. Romoth, Lizzy Gerischer and Mario Stanke*

Bioinformatics 2016 32: 3388-95

Sensitivity (Sn) = TP/(TP+FN), Specificity (Sp) = TN/(TN+FP) at Gene, Exon, or Nucleotide level

		k-way	Gene Sn	Gene Sp	Exon Sn	Exon Sp	Nuc Sn	Nuc Sp
No RNAseq								
D. melanogaster								
AUGUSTUS		-	56.48	60.33	72.88	82.52	92.32	96.90
AUGUSTUScgp		12	66.20	65.79	77.69	85.79	93.57	96.67
N-SCAN$_R$		2	47.91	54.44	68.89	77.20	**94.01**	91.65
CONTRAST		15	**70.49**	**71.99**	**78.07**	**90.76**	91.84	**98.30**
With RNAseq								
AUGUSTUS	*D. mel*		68.01	**73.26**	76.46	**90.20**	93.39	97.57
AUGUSTUScgp	*D. mel*		71.88	72.20	79.27	89.09	96.46	97.49
	D. sim		67.10	67.43	77.64	86.67	95.96	**97.71**
	2 Dros		74.33	73.13	80.22	89.74	96.77	97.57
	4 Dros		**74.46**	73.18	**80.31**	89.79	**96.79**	97.56

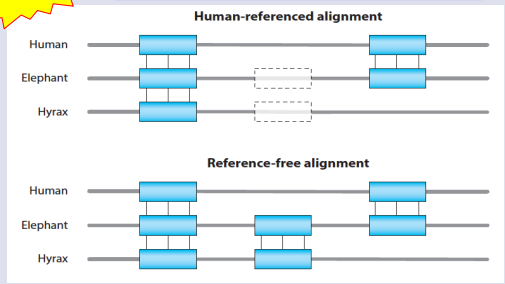
CGP: Comparative Gene Prediction

9

ab initio + similarity methods

Caution!

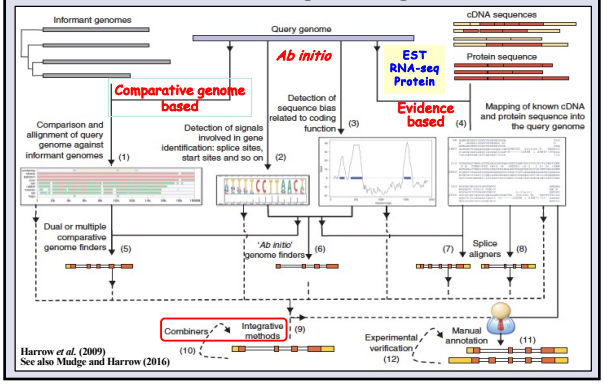
Reference information could generate erroneous insertion or deletion!!



Armstrong *et al.* (2019)

10

Gene-finding strategies

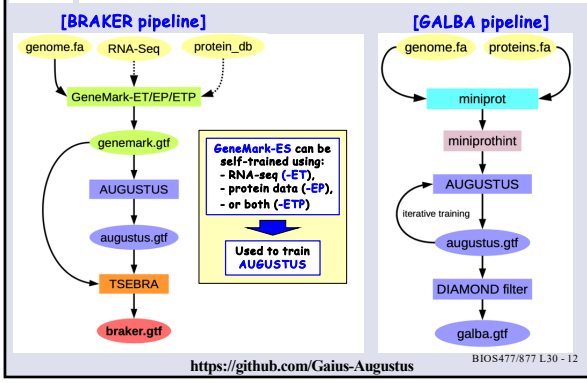


Harrow *et al.* (2009)
See also Mudge and Harrow (2016)

11

Navigating Eukaryotic Genome Annotation Pipelines: A Route Map to BRAKER, Galba, and TSEBRA

Tomáš Brůna[1†], Lars Gabriel[2,3], Katharina J. Hoff[2,3*]



https://github.com/Gaius-Augustus

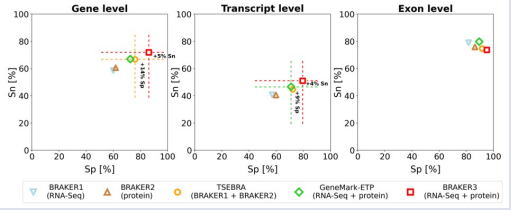
12

BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA

Lars Gabriel [1,2], Tomáš Brůna [3], Katharina J Hoff [1,2], Matthis Ebel [1,2], Alexandre Lomsadze [4], Mark Borodovsky [4,5], Mario Stanke [1,2]

For small genomes (*Drosophila, Arabidopsis, etc.*)

Average accuracy of genome-wide predictions



Sensitivity (Sn) = TP/(TP+FN) [Recall, recovery of actual positives]
Precision = TP/(TP+FP) [Accuracy of predicted positives]

https://github.com/Gaius-Augustus/BRAKER

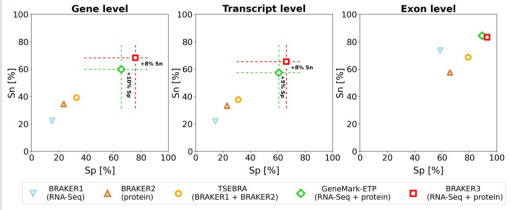
13

BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA

Lars Gabriel [1,2], Tomáš Brůna [3], Katharina J Hoff [1,2], Matthis Ebel [1,2], Alexandre Lomsadze [4], Mark Borodovsky [4,5], Mario Stanke [1,2]

For large genomes (mouse, chicken, *etc.*)

Average accuracy of genome-wide predictions



Sensitivity (Sn) = TP/(TP+FN) [Recall, recovery of actual positives]
Precision = TP/(TP+FP) [Accuracy of predicted positives]

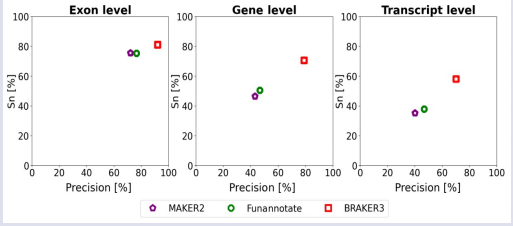
https://github.com/Gaius-Augustus/BRAKER

14

BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS and TSEBRA

Lars Gabriel [1,2], Tomáš Brůna [3], Katharina J Hoff [1,2], Matthis Ebel [1,2], Alexandre Lomsadze [4], Mark Borodovsky [4,5], Mario Stanke [1,2]

Average performance for 11 genomes



Sensitivity (Sn) = TP/(TP+FN) [Recall, recovery of actual positives]
Precision = TP/(TP+FP) [Accuracy of predicted positives]

https://github.com/Gaius-Augustus/BRAKER

15

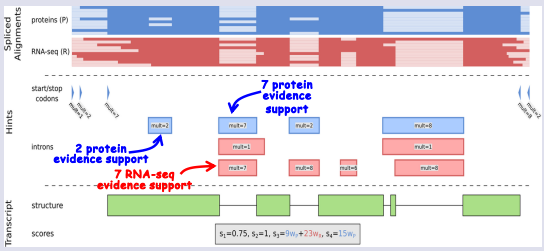
Gabriel et al. BMC Bioinformatics (2021) 22:566
https://doi.org/10.1186/s12859-021-04482-0

BMC Bioinformatics

SOFTWARE Open Access

TSEBRA: transcript selector for BRAKER

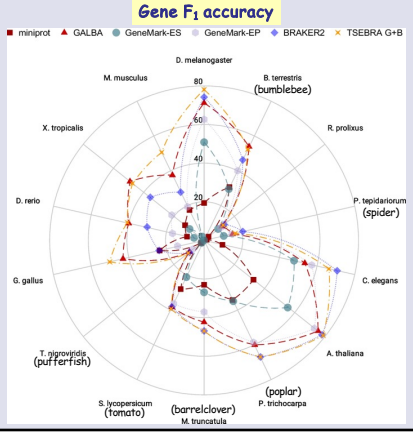
Lars Gabriel[1,2], Katharina J. Hoff[1,2], Tomáš Brůna[3], Mark Borodovsky[4,5] and Mario Stanke[1,2*]



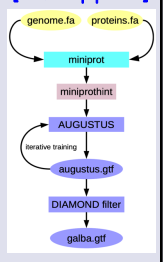
Heterogenious extrinsic evidence (*e.g.*, proteins and RNA-seq) are combined with varied weights:
wP=0.1 for protein hints, wE=10 for RNA-seq hints, wC=5 for protein+RNA-seq hints, and wM=1 for manual hints

16

Gene F_1 accuracy



[GALBA pipeline]

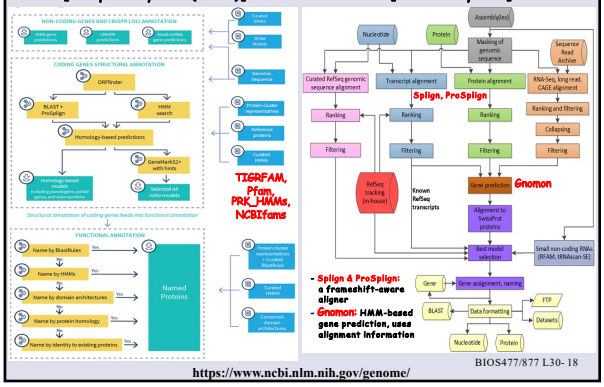


miniprot:
Rapid sliced alignment of protein sequences against the genome

17

NCBI Genome Annotation Pipelines

[For prokaryotes (PGAP)] **[For eukaryotes]**



- **Splign & ProSplign:** a frameshift-aware aligner
- **Gnomon:** HMM-based gene prediction, uses alignment information

https://www.ncbi.nlm.nih.gov/genome/

18

Prokaryotic genome annotation tools

- **PATRIC** (Pathosystems Resource Integration Center) https://www.bv-brc.org
  Assembly, RAST-based annotation, etc.
- **RAST** (Rapid Annotation using Subsystem Technology) https://www.mg-rast.org/
  Gene calling (Glimmer3, GeneMarkS, Prodigal), protein annotation (FIGfams), etc.
- **Prokka** (Rapid prokaryotic genomeannotation)
  https://github.com/tseemann/prokka
  Gene calling (Prodigal), protein annotation using HMMER, etc.
- **IMG/M** (Integrated Microbial Genomes & Microbiomes)
  https://img.jgi.doe.gov/m/
  Assembly, protein-coding gene calling (Prodigal, GeneMark), protein annotation (COGs, Pfam, TIGRFAM, SUPERFAMILY, SMART, CATH-FunFam), etc.
- **KBase** (The DOE Systems Biology Knowledgebase) http://kbase.us/
  Assembly, annotation with RAST, Prokka, many other sequence analyses.

Also in Kimbrel *et al.* (2022) *Methods in Molecular Biology*, vol 2349.

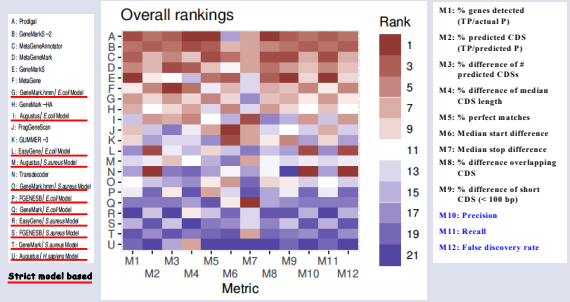
19

No one tool to rule them all: prokaryotic gene prediction tool annotations are highly dependent on the organism of study

Nicholas J Dimonaco, Wayne Aubrey, Kim Kenobi, Amanda Clare, Christopher J Creevey Author Notes

Bioinformatics, Volume 38, Issue 5, 1 March 2022, Pages 1198–1207,

Tested against six genomes: *B. subtilis, C. crescentus, E. coli, M. genitalium, P. fluorescens, and S. aureus*



20

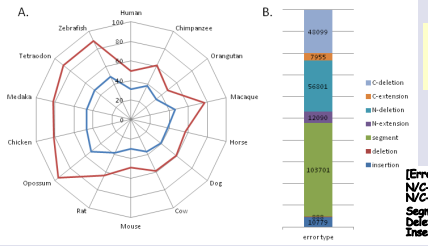
Prosdocimi et al. BMC Genomics 2012, 13:5
http://www.biomedcentral.com/1471-2164/13/5

BMC Genomics

RESEARCH ARTICLE Open Access

Controversies in modern evolutionary biology: the imperative for error detection and quality control

Francisco Prosdocimi[1,2†], Benjamin Linard[3†], Pierre Pontarotti[4], Olivier Poch[3] and Julie D Thompson[3*]



On average 41% of sequences were erroneous

[Error types]
N/C-deletion: N/C-term deletion
N/C-extension: N/C-term extension
Segment: suspicious seq. segment
Deletion: internal deletion
Insertion: internal insertion

Percentage of predicted sequence errors in 19,778 protein families.
Blue: the percentage of sequences with at least one error.
Red: the percentage of total errors observed.

21

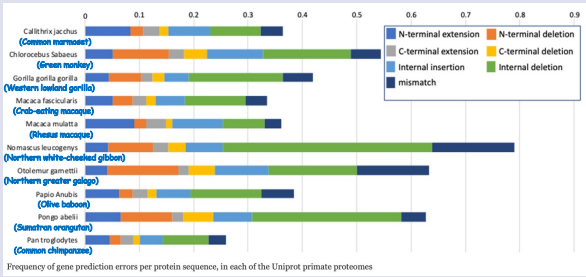
Meyer et al. BMC Bioinformatics (2020) 21:513
https://doi.org/10.1186/s12859-020-03855-1

BMC Bioinformatics

RESEARCH ARTICLE Open Access

Understanding the causes of errors in eukaryotic protein-coding gene prediction: a case study of primate proteomes

Corentin Meyer, Nicolas Scalzitti, Anne Jeannin-Girardon, Pierre Collet, Olivier Poch and Julie D. Thompson*



Frequency of gene prediction errors per protein sequence, in each of the Uniprot primate proteomes

22

Remember…

- Check your data (sequences)!
- Check your alignments!
- Check your phylogenies!
- Check your outputs!!!

23

Assignment #12: due May 13

Final Graduate Report: due May 14

Course Evaluation on Canvas
Complete by
11:59 PM, May 11 (Sat)

24