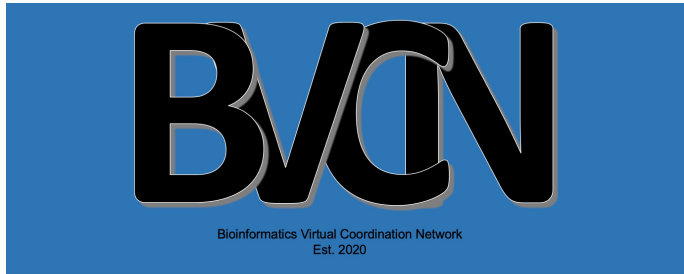
Functional Annotation Lesson 1

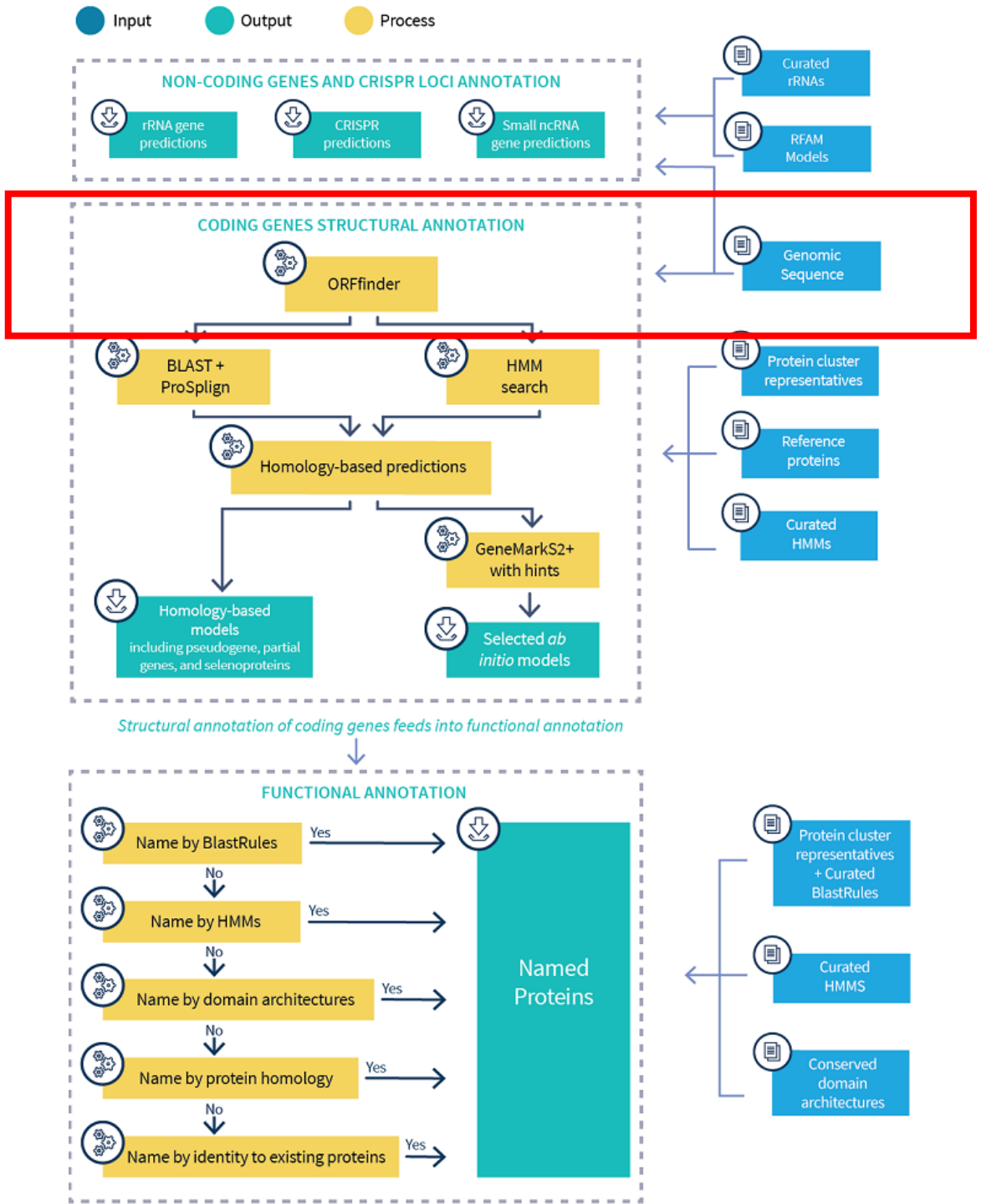
Predicting proteins

- Open reading frames in Bacteria and Archaea
- Open reading frames in Eukaryotes
- Translation tables

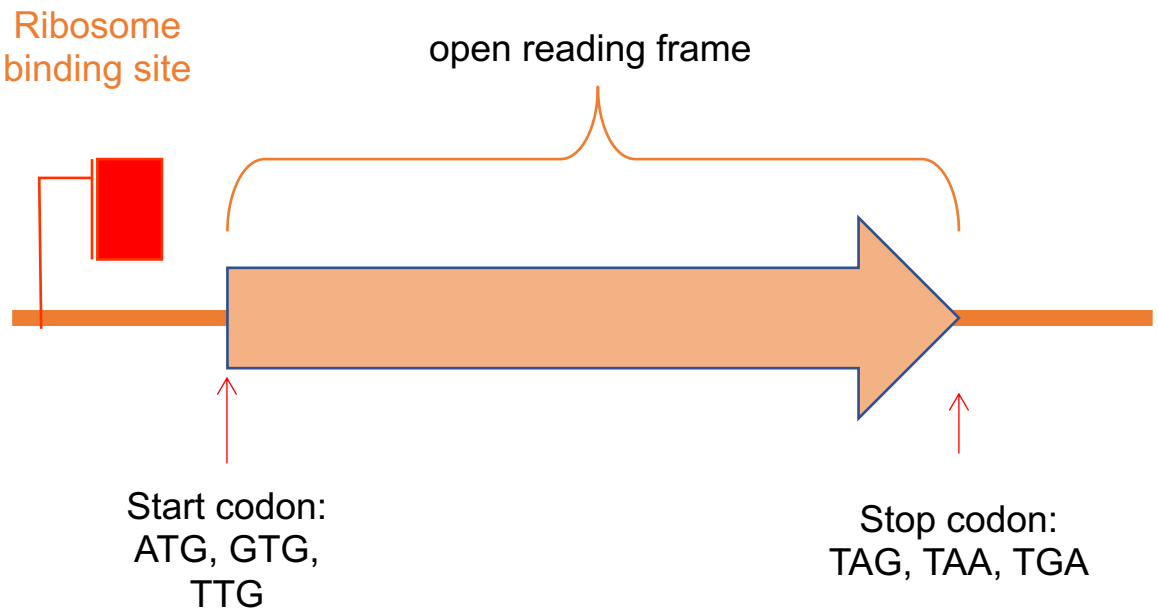


NCBI PGAP

Prokaryotic Genome Annotation Pipeline



Bacteria & Archaea Gene Prediction



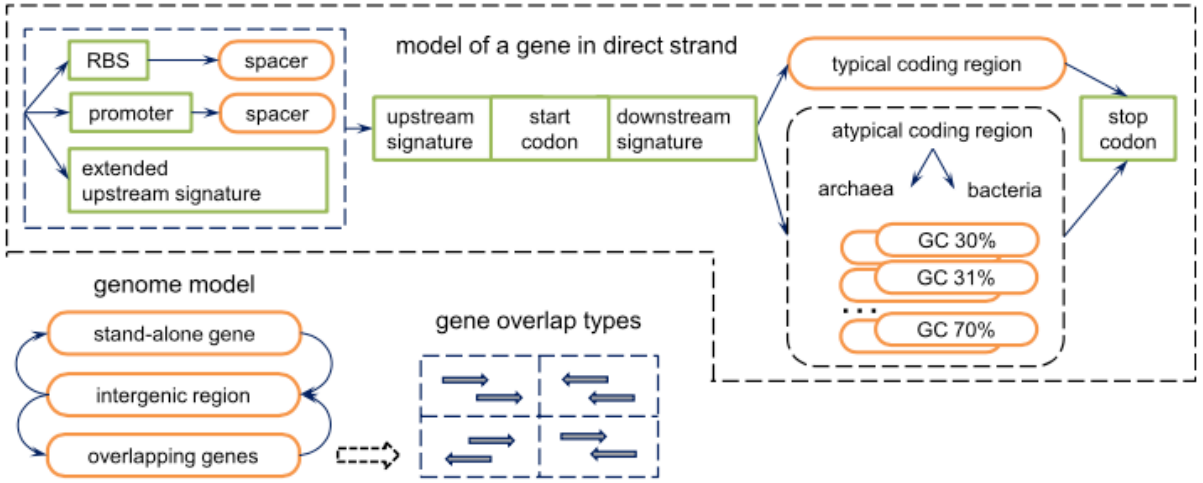
Gene model used by all *ab initio* gene finder

Traits a gene finder is looking for:

- Ribosome-binding site within a predicted distance from a start codon
- One of 3 start codons
- One of 3 stop codons
- No frame interruptions
  - Most gene finders do not handle point mutations that generate missense proteins or superfluous stop codons

Common Gene Prediction Tools

GeneMarkS-2 (2018)



Missed annotated genes (FN)

GeneMarkS		136	**494**	434	192	296	1552
Glimmer3		**66**	678	1170	341	323	2578
Prodigal		161	639	417	92	78	1387
GeneMarkS-2		132	596	**370**	**76**	**69**	**1243**
B	**Bins (nt)**	**<150**	**150–300**	**300–600**	**600–900**	**>900**	**Total**

Currently used by NCBI PGAP

Prodigal (2012)

- Predicts protein-coding genes
- Handles draft genomes and metagenomes
- Runs quickly
- Runs unsupervised (unsupervised ML algorithm)
- Handles gaps, scaffolds, and partial genes
- Identifies translation initiation sites
- Outputs detailed summary statistics for each genome

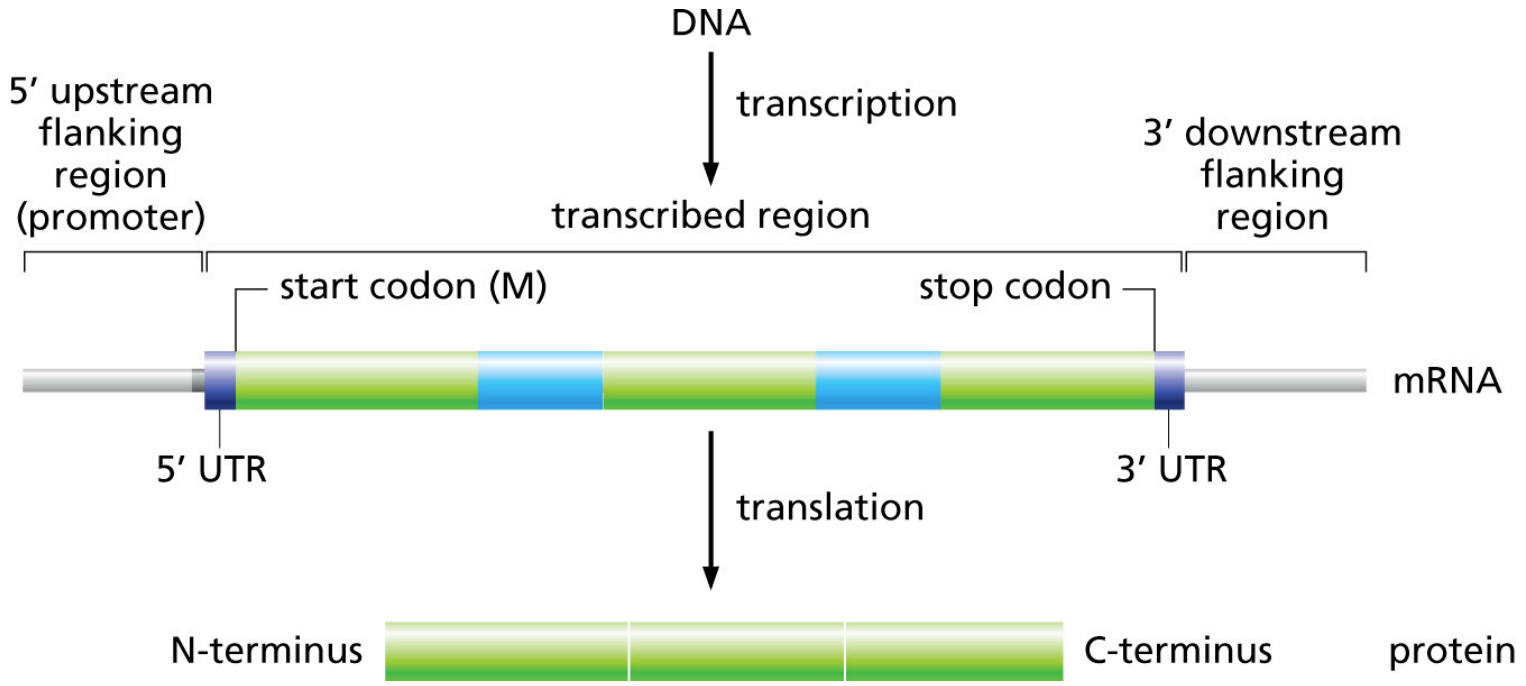
Currently used by JGI

Lomsadze (2018)

Eukaryotic Gene Prediction

Eukaryotic gene prediction requires consideration of:

- Introns
- Exons
- Untranslated regions (UTRs)
- Splice forms



http://www.garlandscience.com/textbooks/0815340249.asp?type=artwork

Common Gene Prediction Tools

Braker2 (2018)

Preparation steps:

- Create a repeat library and mask your genome with RepeatMasker
- Trim and align RNA-seq to genome
- Align transcripts, ESTs, or long transcript to genome with splice aware aligner
- RUN Braker2

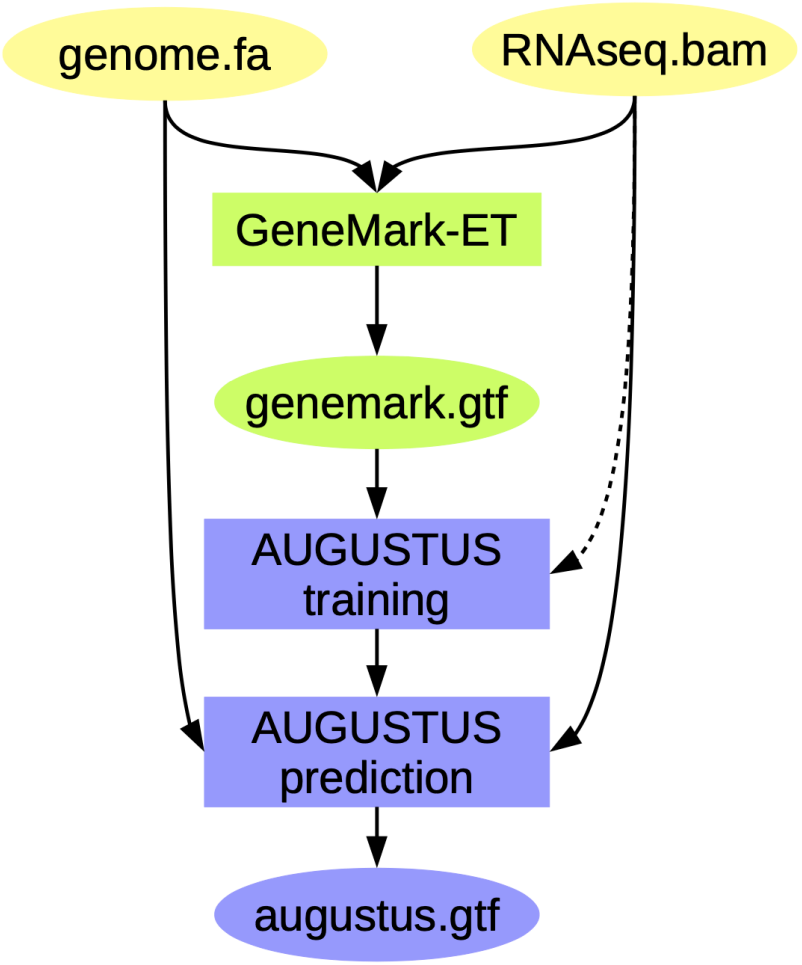
trimgalore

Hisat2

Picard

BRAKER1, a combination of GeneMark-ET and AUGUSTUS, that uses genomic and RNA-Seq data to automatically generate full gene structure annotations in novel genome

BRAKER2 is an extension of BRAKER1. BRAKER2 reaches high gene prediction accuracy even in the absence of the annotation of very closely related species and in the absence of RNA-Seq data.



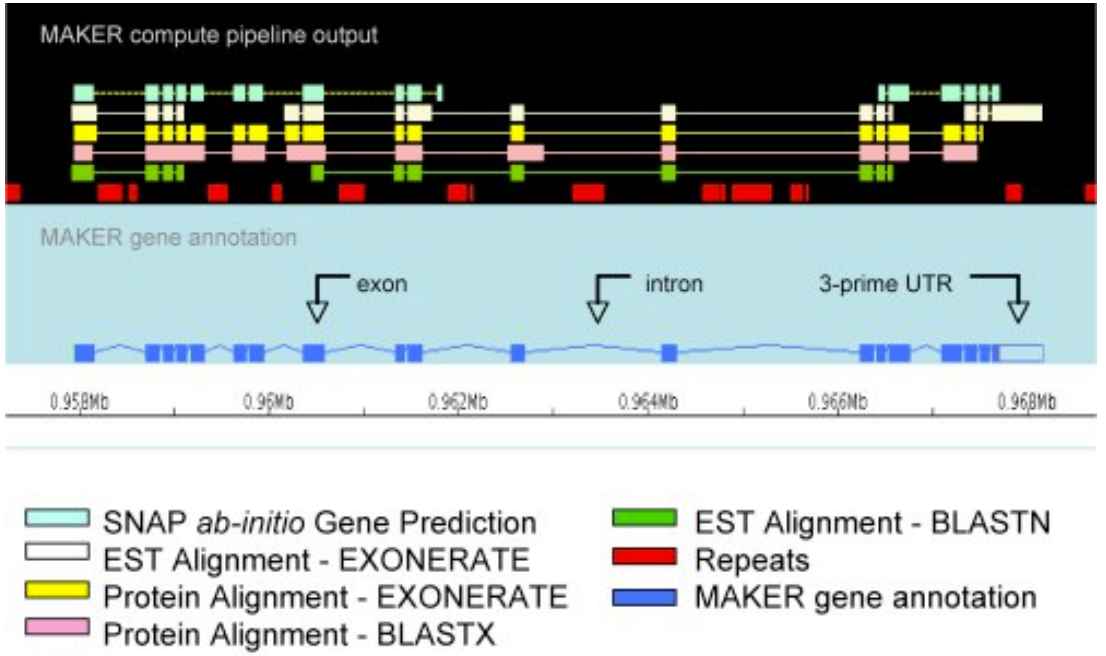
Common Gene Prediction Tools

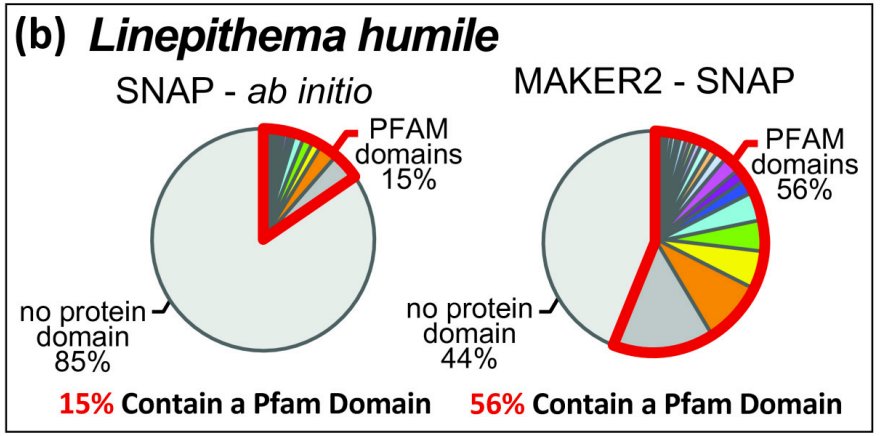
MAKER (v3 2020)

Steps:

- Mask out repeats
- Align EST to genome
- Align proteins to genome
- Produce ab initio predictions
- Synthesize into final predictions

MAKER is an annotation pipeline, not a gene predictor. MAKER does not predict genes, rather MAKER leverages existing software tools (some of which are gene predictors) and integrates their output to produce what MAKER finds to be the best possible gene model for a given location based on evidence alignments.





Common Gene Prediction Tools

Bacteria and Archaea

ab initio prediction → uses a set of basic rules to determine open reading frames

- The more “standard” the genome, the better the tool works
- Possible to design rules for alternative genotypes

Eukaryotes

Reliant on previous data sets

- Transcriptomic
- Close evolutionary relative
- Conglomeration of multiple tools and datasets leveraging as much data as possible

Translation tables

- 1. The Standard Code
- 2. The Vertebrate Mitochondrial Code
- 3. The Yeast Mitochondrial Code
- 4. The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code
- 5. The Invertebrate Mitochondrial Code
- 6. The Ciliate, Dasycladacean and Hexamita Nuclear Code
- 9. The Echinoderm and Flatworm Mitochondrial Code
- 10. The Euplotid Nuclear Code
- 11. The Bacterial, Archaeal and Plant Plastid Code
- 12. The Alternative Yeast Nuclear Code
- 13. The Ascidian Mitochondrial Code
- 14. The Alternative Flatworm Mitochondrial Code
- 16. Chlorophycean Mitochondrial Code
- 21. Trematode Mitochondrial Code
- 22. Scenedesmus obliquus Mitochondrial Code
- 23. Thraustochytrium Mitochondrial Code
- 24. Pterobranchia Mitochondrial Code
- 25. Candidate Division SR1 and Gracilibacteria Code
- 26. Pachysolen tannophilus Nuclear Code
- 27. Karyorelict Nuclear Code
- 28. Condylostoma Nuclear Code
- 29. Mesodinium Nuclear Code
- 30. Peritrich Nuclear Code
- 31. Blastocrithidia Nuclear Code
- 33. Cephalodiscidae Mitochondrial UAA-Tyr Code

33 translation tables – organism dependent

https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi

Lesson 2

Predicting function using orthologous comparisons