Introduction to *ab initio* and evidence-based gene finding

Wilson Leung 12/2023

1

Outline

- Overview of computational gene predictions
- Different types of eukaryotic gene predictors
- Common types of gene prediction errors

2

Computational gene predictions

- Identify genes within genomic sequences
  - **Protein-coding genes**
  - Non-coding RNA genes
  - Regulatory regions (enhancers, promoters)
- Predictions must be confirmed experimentally
  - Eukaryotic gene predictions have **high error rates**
  - Two major types of RefSeq records:
    - NM_/NP_ = experimentally confirmed
    - XM_/XP_ = computational predictions

3

Primary goal of computational gene prediction algorithms

- Label each nucleotide in a genomic sequence
  - Identify the **most likely sequence of labels** (i.e., optimal path)

4

Basic properties of gene prediction algorithms

- Model must satisfy **biological constraints**
  - Coding region must begin with a start codon
  - Initial exon must occur before splice sites and introns
  - Coding region must end with a stop codon
- Model rules using a finite state machine (FSM)
- Use **species-specific characteristics** to improve the accuracy of gene predictions
  - Distribution of exon and intron sizes
  - Base frequencies (e.g., GC content, codon bias)
  - Protein sequences from the same or closely related species

5

Prokaryotic gene predictions

- Prokaryotes have relatively simple gene structure
  - Single open reading frame
  - Alternative start codons: AUG, GUG, UUG
- Gene finders can predict most prokaryotic genes accurately (> 90% sensitivity and specificity)
  - Glimmer
    - Salzberg S., et al. *Microbial gene identification using interpolated Markov models*, NAR. (1998) **26**, 544-548
  - NCBI Prokaryotic Genome Annotation Pipeline (PGAP)
    - Li W., et al. *RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation*, NAR. (2021) **49**(D1), D1020-D1028
    - https://github.com/ncbi/pgap

6

Eukaryotic gene predictions have high error rates

- Gene finders generally do **a poor job (<60%)** predicting genes in eukaryotes
- More **variations** in the gene models
  - Alternative splicing (multiple isoforms)
  - Non-canonical splice sites (e.g., *toy*)
  - Non-canonical start codon (e.g., *Fmr1*)
  - Stop codon read through (e.g., *gish*)
  - Nested genes (e.g., *ko*)
  - Trans-splicing (e.g., *mod(mdg4)*)
  - Pseudogenes (e.g., *swaPsi*)

7

Types of eukaryotic gene predictors

- *Ab initio*
  - **GENSCAN**, **geneid**, **SNAP**, **GlimmerHMM**
- Evidence-based (extrinsic)
  - **Augustus**, **genBlastG**, **GeMoMa**, Exonerate, GenomeScan
- Comparative genomics
  - **Twinscan/N-SCAN**, **SGP2**
- Transcriptome-based (RNA-Seq)
  - **Cufflinks**, **StringTie**, **Trinity**, CodingQuarry
- Combine *ab initio* and evidence-based approaches
  - **Gnomon**, MAKER, EASEL, EVM, JIGSAW, IPred, GLEAN

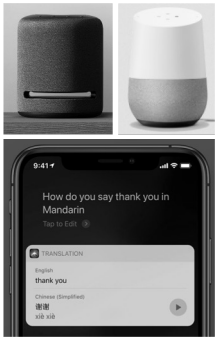
8

Ab initio gene prediction

- *Ab initio* = from the beginning
- Predict genes using only the genomic DNA sequence
  - Search for **signals** of protein coding regions
  - Based on a probabilistic model
    - Hidden Markov Models (HMM)
    - Support Vector Machines (SVM)
- GENSCAN
  - Burge C. and Karlin S. *Prediction of complete gene structures in human genomic DNA*, JMB. (1997) **268**, 78-94

9

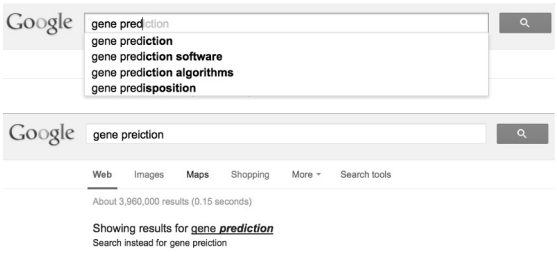
Hidden Markov Models (HMM)



- A type of supervised machine learning algorithm
  - Uses **Bayesian statistics**
  - Makes classifications based on characteristics of **training data**
- Many types of applications
  - Speech and gesture recognition
  - Bioinformatics
    - Gene predictions
    - Sequence alignments
    - ChIP-seq analysis
    - Protein folding

10

Supervised machine learning

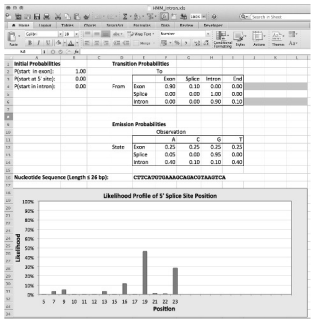


Use previous search results to predict search terms and correct spelling errors

Norvig P. How to write a spelling corrector. https://www.norvig.com/spell-correct.html

11

GEP curriculum on HMM



- Use an HMM to predict a splice donor site
  - Use Excel to experiment with different emission and transition probabilities
- See the **Curriculum** section of the GEP website
  - Also available on CourseSource

Weisstein AE et al. *A Hands-on Introduction to Hidden Markov Models*. CourseSource. (2016), https://doi.org/10.24918/cs.2016.8

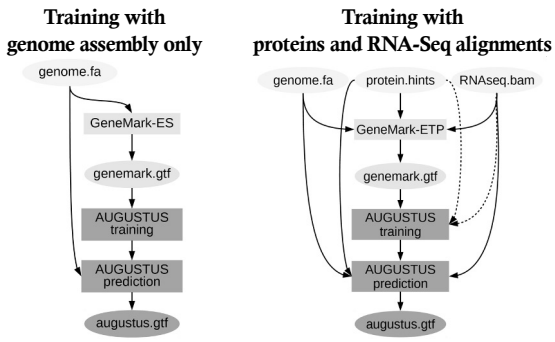
12

Ways to create **training sets** to estimate transition and emission parameters

- Manually curated genes for the target species
- Bootstrap with *ab initio* gene predictions
  - GeneMark-ES, GENSCAN
- Sequence similarity to orthologs in informant species
  - BUSCO, BRAKER2
- Whole genome conservation profiles
  - Augustus-cgp, N-SCAN, SGP2
- RNA-Seq (splice junctions, assembled transcripts)
  - BRAKER1

13

BRAKER2 training protocols

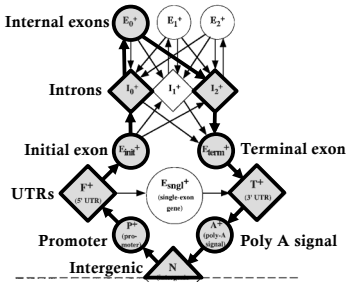


Hoff KJ. *BRAKER 2 User Guide*. https://github.com/Gaius-Augustus/BRAKER

14

GENSCAN HMM Model

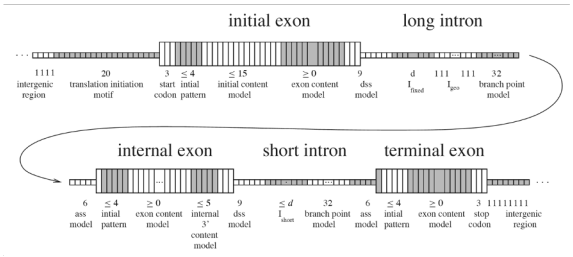
- GENSCAN considers:
  - Promoter, splice sites and polyadenylation signals
  - Hexamer frequencies and base compositions
  - Probability of coding and non-coding DNA
  - Distributions of gene, exon and intron lengths



Burge C. and Karlin S. *Prediction of complete gene structures in human genomic DNA*, JMB. (1997) **268**, 78-94

15

Use multiple HMMs to describe different parts of a gene

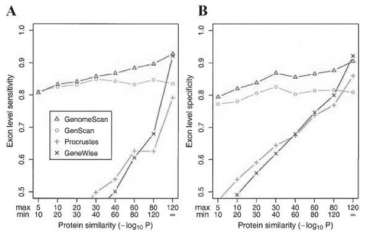


Stanke M. and Waack S. *Gene prediction with a hidden Markov model and a new intron submodel.* Bioinformatics. (2003) **19**, Suppl 2:ii215-25.

16

Evidence-based gene predictions

- Use sequence alignments to improve predictions
  - EST, cDNA or protein from closely-related species



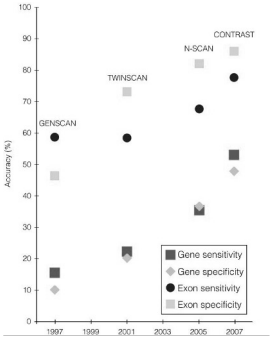
Exon sensitivity: Percent of real exons identified

Exon specificity: Percent of predicted exons that are correct

Yeh RF, et al. *Computational Inference of Homologous Gene Structures in the Human Genome*, Genome Res. (2001) **11**, 803-816

17

Predictions using comparative genomics

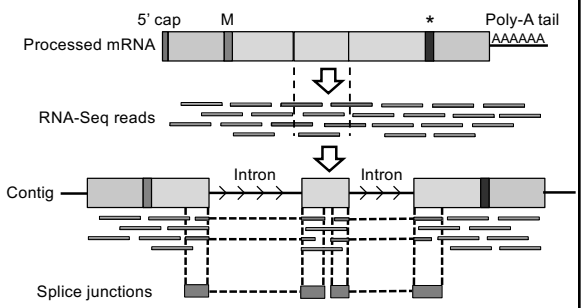


- Use whole genome alignments from one or more **informant species**
- CONTRAST predicts 50% of genes correctly
- Requires **high quality whole genome alignments and training data**

Flicek P. *Gene prediction: compare and CONTRAST*. Genome Biology (2007) **8**, 233

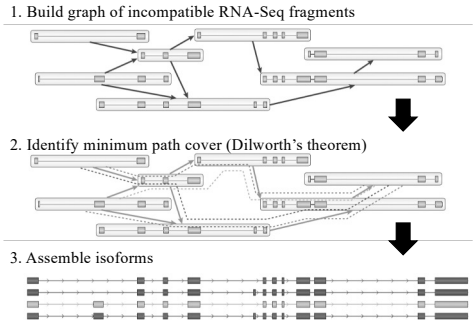
18

Intron predictions based on spliced RNA-Seq reads



19

Cufflinks – reference-based transcriptome assembly

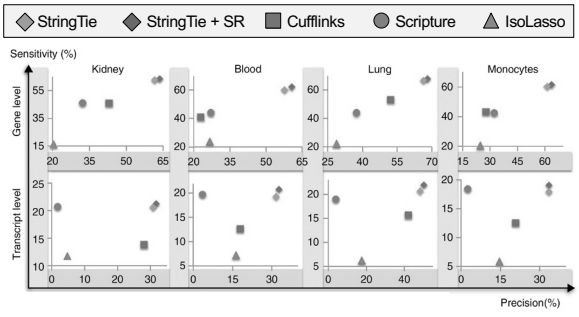


- Use TransDecoder to identify coding regions within assembled transcripts

Martin JA, Wang Z. *Next-generation transcriptome assembly*. Nat Rev Genet. (2011) **12(10),** 671-82

20

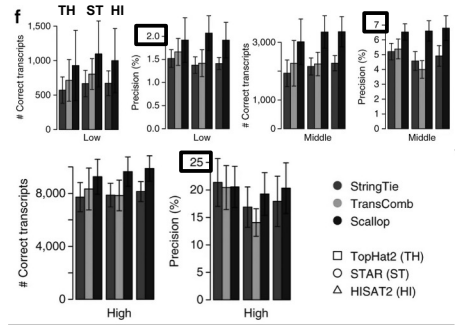
Transcriptome assembly of RNA-Seq reads remains an area of **active research**



Pertea M et al. Nat Biotechnol. 2015 33(3):290-5.

21

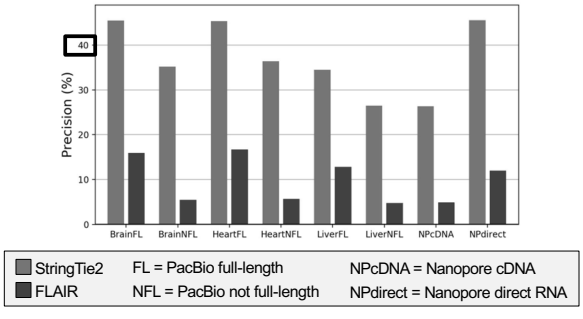
Precision of assembling Illumina RNA-Seq reads into transcripts depends on expression levels



Shao M and Kingsford C. Nat Biotechnol. 2017 Dec;35(12):1167-1169.

22

Improve transcript assembly using PacBio IsoSeq and Nanopore RNA-Seq



Kovaka S et al. Genome Biol. 2019 Dec 16;20(1):278.

23

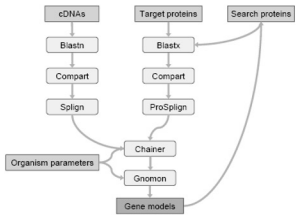
Generate consensus gene models

- Gene predictors have different strengths and weaknesses
- Create **consensus gene models** by combining results from multiple gene finders and sequence alignments
  - EVidenceModeler (EVM)
    - Haas BJ et al. *Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments*. Genome Biology (2008) **9**(1), R7
  - TSEBRA
    - Gabriel L et al. *TSEBRA: transcript selector for BRAKER*. BMC Bioinformatics (2021) **22**(1), 566

24

Automated annotation pipelines

NCBI Gnomon gene prediction pipeline



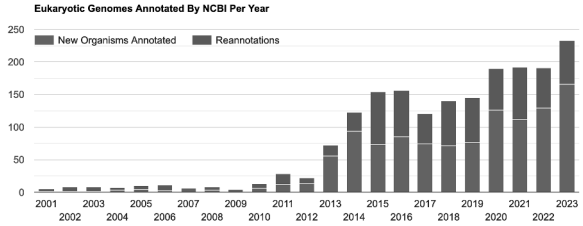
- Integrate **biological evidence** into the predicted gene models
- Examples:
  - NCBI Gnomon
  - Ensembl
  - UCSC Gene Build
- EGASP results for the Ensembl pipeline:
  - 71.6% gene sensitivity
  - 67.3% gene specificity

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

25

Eukaryotic genomes annotated by NCBI

- RefSeq annotations available for more than **1,100 species**

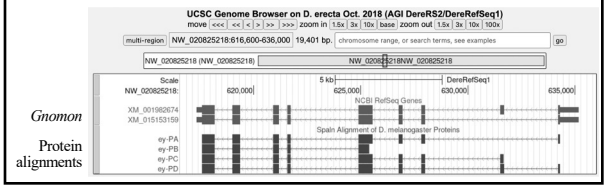


https://www.ncbi.nlm.nih.gov/genome/annotation_euk/#graphs

26

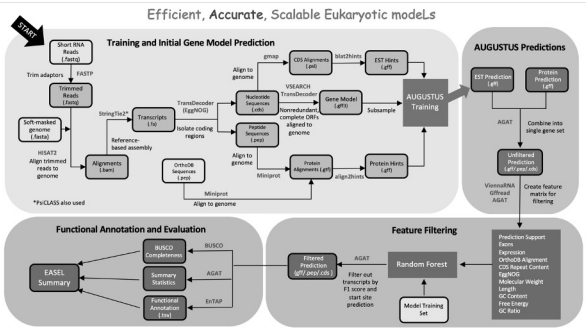
Drosophila RefSeq gene predictions

- Based on **RNA-Seq** data from either the same or closely-related species
  - Predictions include **untranslated regions** and **multiple isoforms**
- Gnomon gene predictions are available through the NCBI RefSeq database:
  - https://www.ncbi.nlm.nih.gov/genome/annotation_euk/all/



27

EASEL — incorporate **feature filtering** and **functional annotations** into the workflow

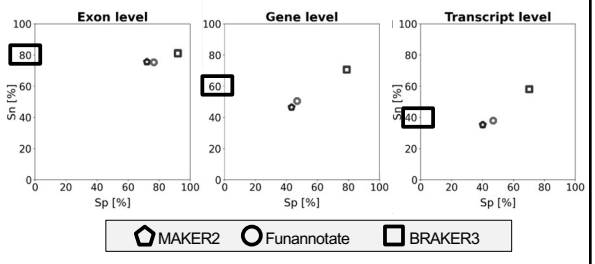


https://gitlab.com/PlantGenomicsLab/easel

28

Most genomes can benefit from **manual gene annotations**

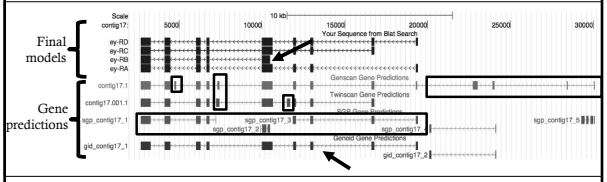
- Average sensitivity (Sn) and sensitivity (Sp) of gene predictions for eight eukaryotic genomes



Gabriel L et al. bioRxiv 2023.06.10.544449; doi: 10.1101/2023.06.10.544449

29

Common problems with gene finders



- Split single gene into multiple predictions
- Fused with neighboring genes
- Missing exons
- Over predict exons or genes
- Missing isoforms

30

Non-canonical splice donors and acceptors

- Many gene predictors strongly prefer models with canonical splice donor (**GT**) and acceptor (**AG**) sites
- Check **Gene Record Finder** or FlyBase for genes that use non-canonical splice sites in *D. melanogaster*

Introns with Non-canonical Splice Sites

Transcript Name	FlyBase ID	Splice Donor	Splice Acceptor
Cadps-RD	intron_Cadps:22_Cadps:23	GC	AG

Frequency of non-canonical splice sites in FlyBase Release 6.55 (Number of unique introns: 72,061)

Donor site	Count
GC	603
AT	30
GA	15

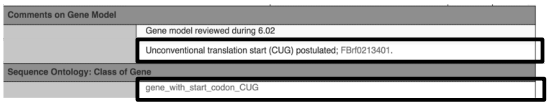
Acceptor site	Count
AC	34
TG	28
AT	16

31

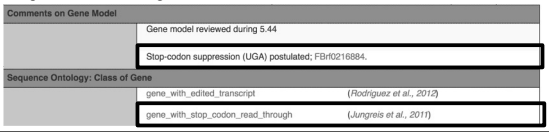
Annotate unusual features in gene models using *D. melanogaster* as a reference

- Examine the "**Comments on Gene Model**" and the "**Sequence Ontology**" sections of the FlyBase Gene Report

Non-canonical start codon:

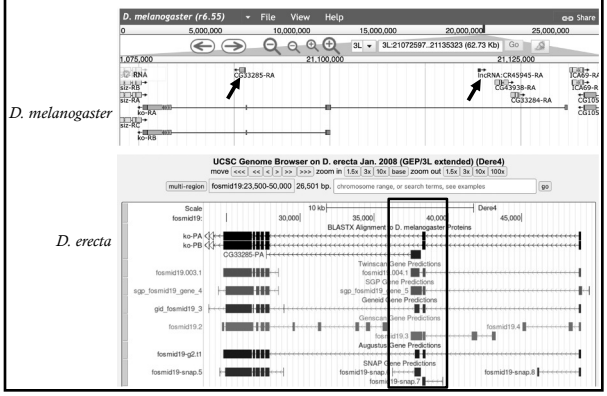


Stop codon read through:



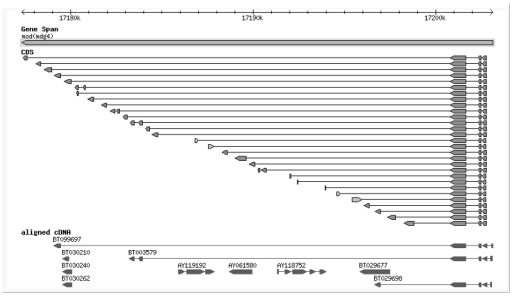
32

Nested genes in *Drosophila*



33

Trans-spliced gene in *Drosophila*



A special type of RNA processing where exons from two primary transcripts are ligated together

34

Gene prediction results for the GEP annotation projects

- Gene prediction results are available through the GEP UCSC Genome Browser mirror
  - Under the **Genes and Gene Prediction Tracks** section
- Access the predicted peptide sequence:
  - Click on the feature, and then click on the **Predicted Protein** link

35

Summary

- Gene predictors can quickly identify potentially interesting features within a genomic sequence
- The predictions are hypotheses that must be confirmed experimentally
- Eukaryotic gene predictors generally can accurately identify internal exons
  - Much lower sensitivity and specificity when predicting complete gene models

36