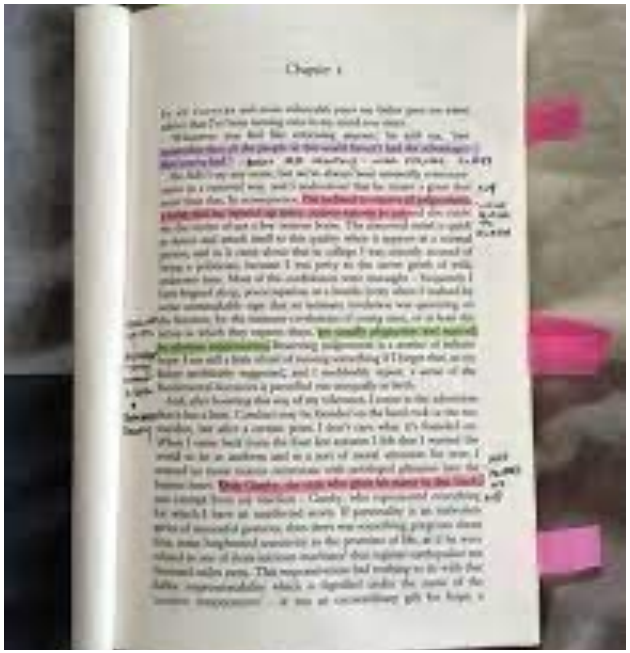
Intro to Genome Annotation

Carlos F. Arias M

SMSC 2023

An-no-ta-tion \ ˌ a-nə-ˈtā-shən\

- **A critical or explanatory note or body of notes added to a text.**
- **The act of annotating.**



Genome Anotation

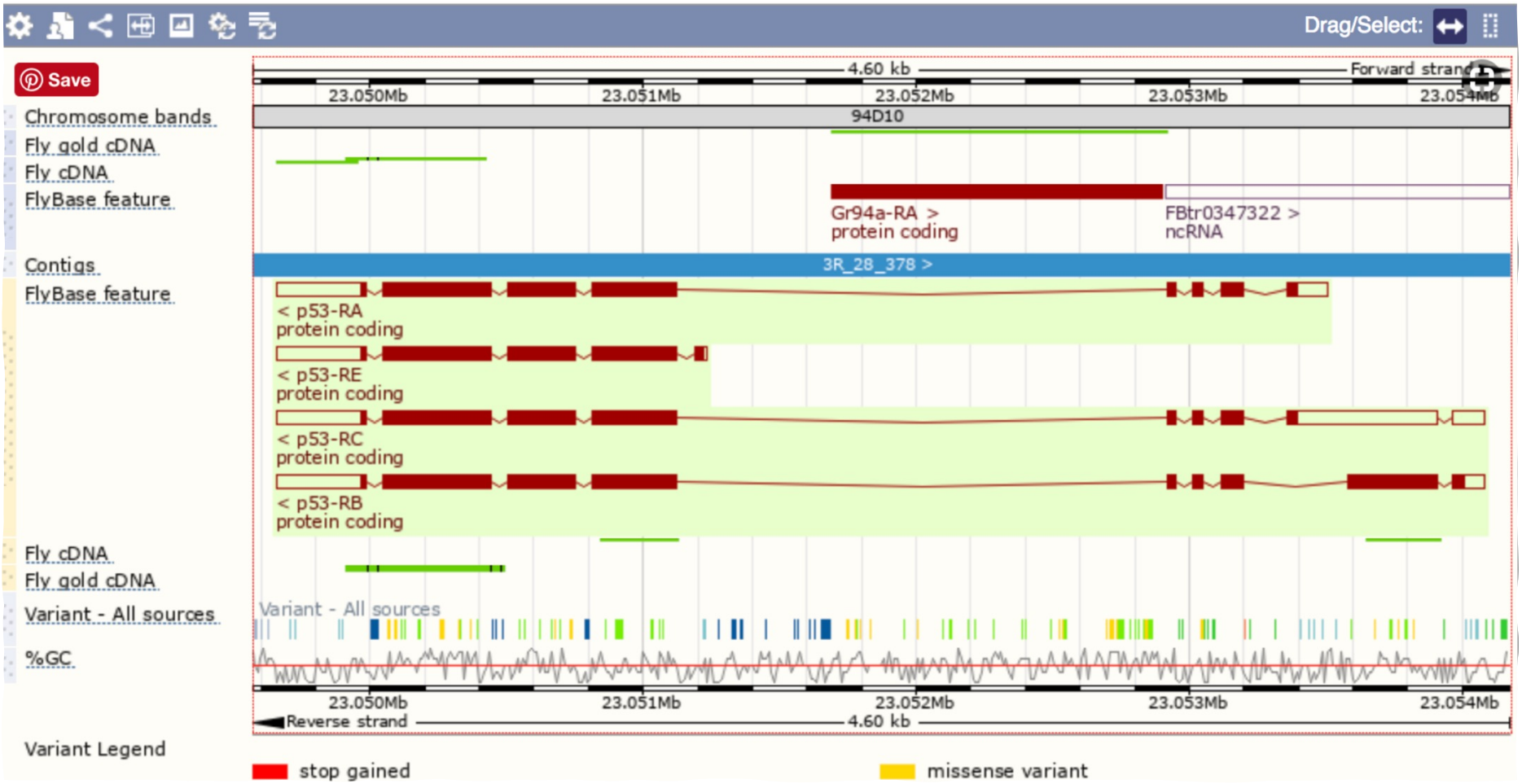
is the process of identifying different elements in a genome assembly

Two steps in genome annotation

- Identifiying were are the genes on the genome (i.e structural annotation).

Chromosome 3R: 23,049,569-23,054,170

Chromosome 3R: 23,049,569-23,054,170



Finding genes in a see of nucleotides

Two approaches:

Homology-based gene prediction

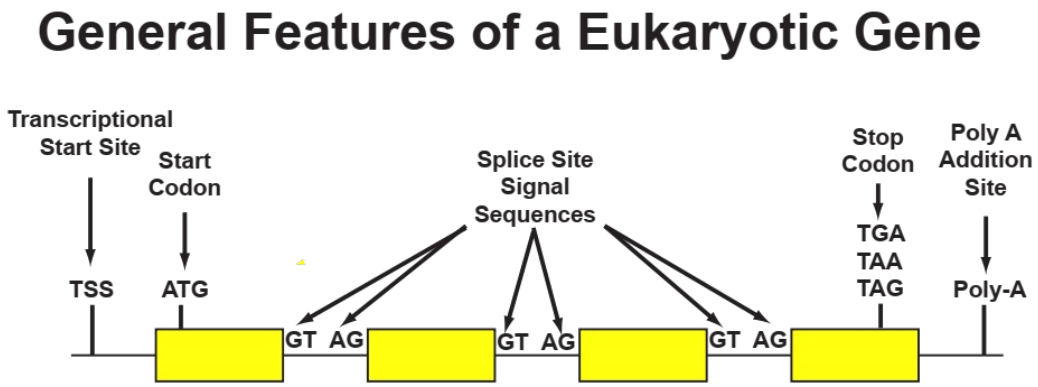
- Similarity Searches (*e.g.* BLAST, BLAT)
- Genome Browsers
- RNA evidence (ESTs)

Ab initio gene prediction

- Gene prediction programs
- Prokaryotes
  - ORF identification
- Eukaryotes
  - Promoter prediction
  - PolyA-signal prediction
  - Splice site, start/stop-codon predictions

Ab initio gene prediction

- Ab inition: “From firs principals”
- Requires only a genomic sequence.
- Uses statistical model of genome composition to identify most probable location of start/stop codons, splice sites.
- Popular implementations
  - Augustus
  - SNAP
  - GeneMark



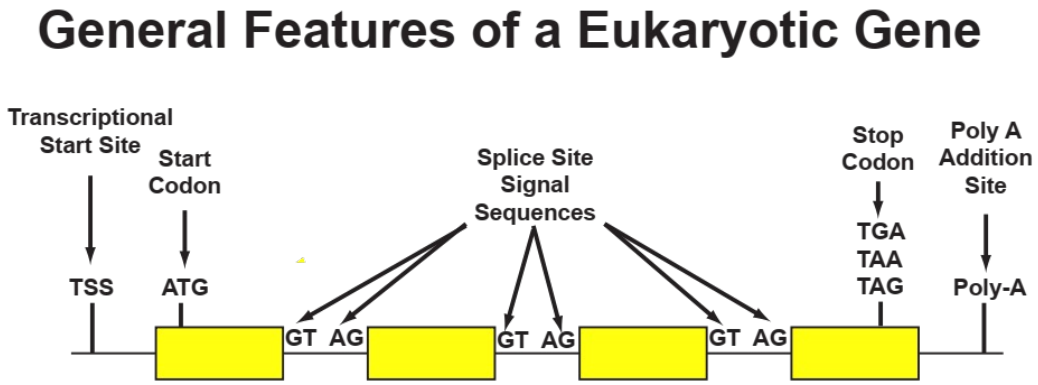
Ab initio gene prediction

- Rule-based programs

  Use explicit set of rules to make decisions
- Neural Network-based programs

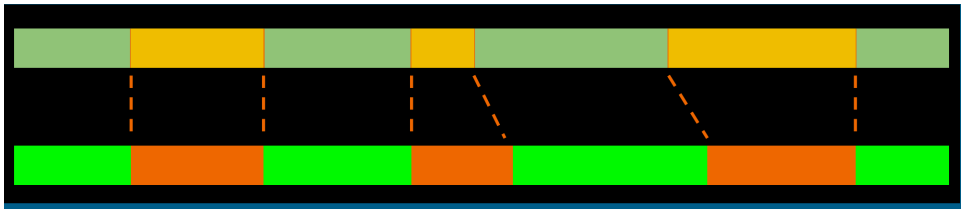
  Use dataset to build rules
- Hidden Markov Model-based programs

  Use probabilities of states and transitions between these states to predict features.



Homology-based gene prediction

- Utilizes experimental (transcript) and/or homology (reference proteins) data .
- Spliced alignment of sequences reveals gene structure.
  - matches = exons
  - gaps = introns
- Requires annotated genomes of closely related taxa
- Popular implementations
  - GeMoMa
  - Exonerate



Comparison of prediction methods

Ab Inition	Homology-Base
Do not require extrinsic evidence	Requires transcript and/or protein sequences
Does not benefit from additional transcript data	Accuracy improves with additional transcript data
More likely to recover complete gene structures	More likely to recover accurate internal exon/intron structure

Our Genome annotation

Two Steps

- **Repeat identification and masking (with Repeatmodeler and RepeatMasker**
- **Homology- Based Annotation with GeMoMa pipeline**

GeMoMa Pipeline

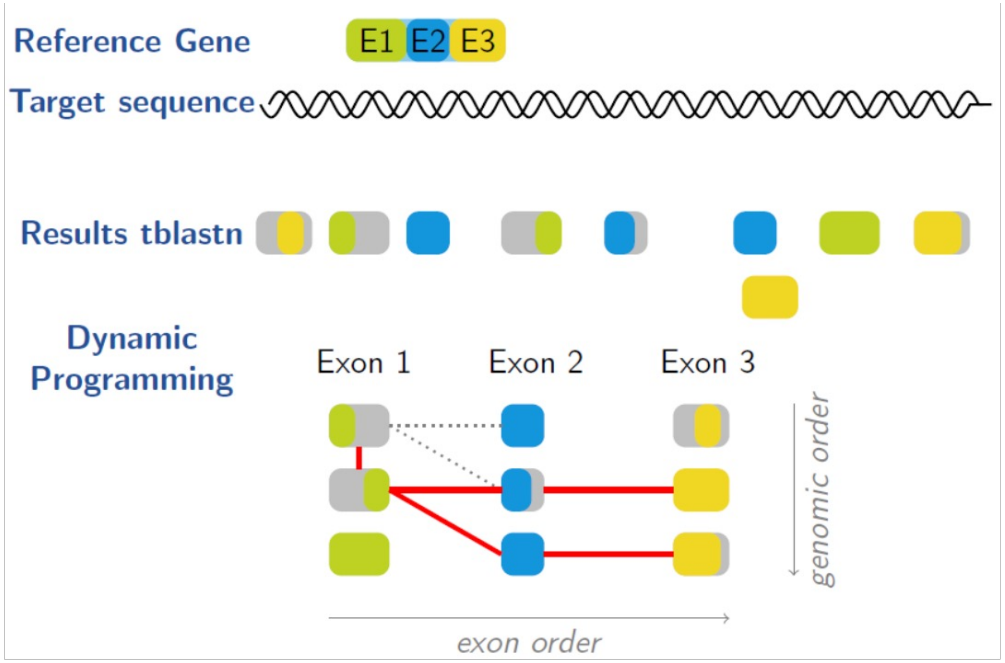


Figure 1: Illustration of the GeMoMa algorithm. GeMoMa uses `tblastn` to search for homologs of all (partially) coding exons of the reference transcript. Subsequently, a dynamic programming algorithm is used to determine the best combination of the hits.

GeMoMa Pipeline

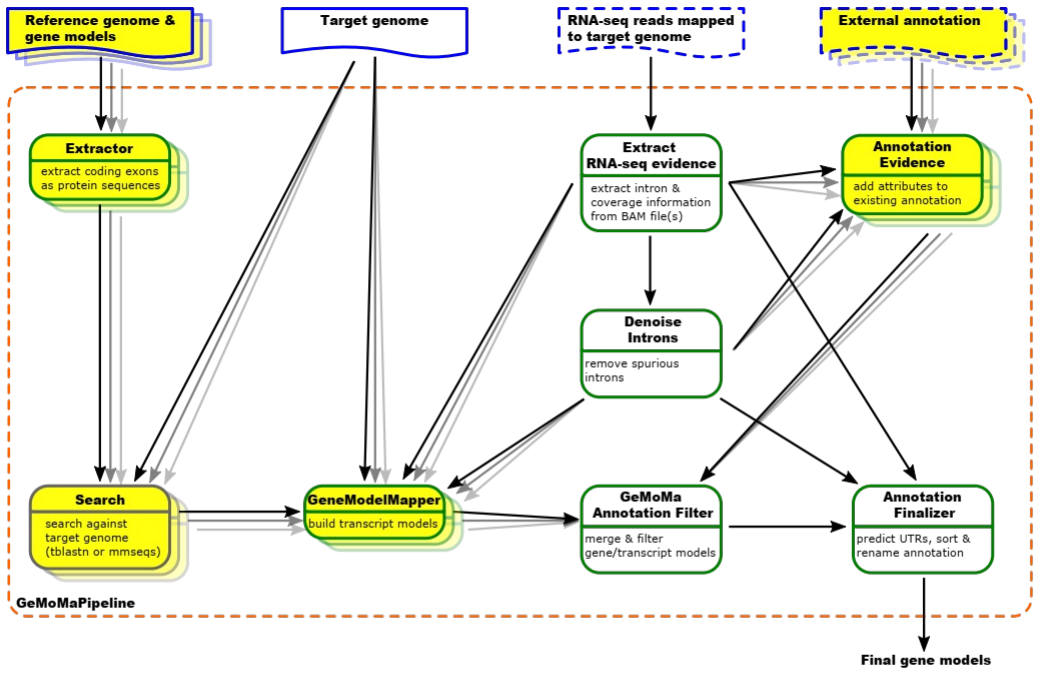


Figure 2: GeMoMa workflow. Solid blue items represent input data sets, dashed blue items are optional inputs, green boxes represent GeMoMa modules, while grey boxes represent external modules. The GeMoMa Annotation Filter allows to combine predictions from different reference species. RNA-seq data is optional.