Genome Annotation

(adapted from A. Bombarely IBMCP)

Ana Conesa, Professor

Institute for Integrative Systems Biology , CSIC, Spain

ana.conesa@csic.es @anaconesa @conesa_lab

MINISTERIO
DE CIENCIA, INNOVACIÓN
Y UNIVERSIDADES

CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

CSIC HUB
BIOLOGÍA
COMPUTACIONAL Y
BIOINFORMÁTICA (BCB)

Who I am?

Montevideo
Valencia

Who I am?

Who I am?

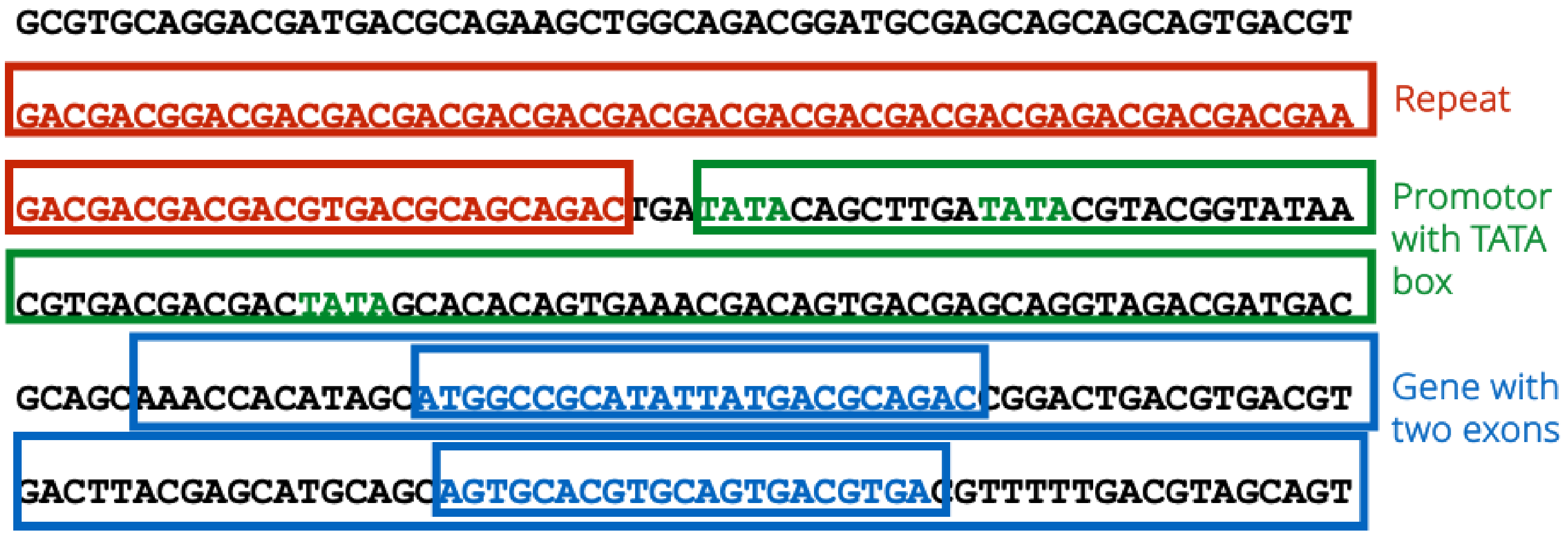
My hobby: my cats

Let´s get serious... genome annotation....

```
GCGTGCAGGACGATGACGCAGAAGCTGGCAGACGGATGCGAGCAGCAGCAGTGACGT

GACGACGGACGACGACGACGACGACGACGACGACGACGACGAGACGACGACGAA

GACGACGACGACGTGACGCAGCAGACTGATATACAGCTTGATATACGTACGGTATAA

CGTGACGACGACTATAGCACACAGTGAAACGACAGTGACGAGCAGGTAGACGATGAC

GCAGCAAACCACATAGCATGGCCGCATATTATGACGCAGACCGGACTGACGTGACGT

GACTTACGAGCATGCAGCAGTGCACGTGCAGTGACGTGACGTTTTTGACGTAGCAGT
```

Do all nucleotides have the same function?

Genome annotation



GCGTGCAGGACGATGACGCAGAAGCTGGCAGACGGATGCGAGCAGCAGCAGTGACGT

GACGACGGACGACGACGACGACGACGACGACGACGACGACGACGAGACGACGACGAA

GACGACGACGACGTGACGCAGCAGACTGATATACAGCTTGATATACGTACGGTATAA

CGTGACGACGACTATAGCACACAGTGAAACGACAGTGACGAGCAGGTAGACGATGAC

GCAGCAAACCACATAGCATGGCCGCATATTATGACGCAGACCGGACTGACGTGACGT

GACTTACGAGCATGCAGCAGTGCACGTGCAGTGACGTGACGTTTTTGACGTAGCAGT

Repeat

Promotor with TATA box

Gene with two exons

Genome annotation is about identifying functional elements in a DNA sequence

Genome elements

Functional

Genome structure

- ▤ Telomeres
- ● Centromeres

Expression regulatory elements

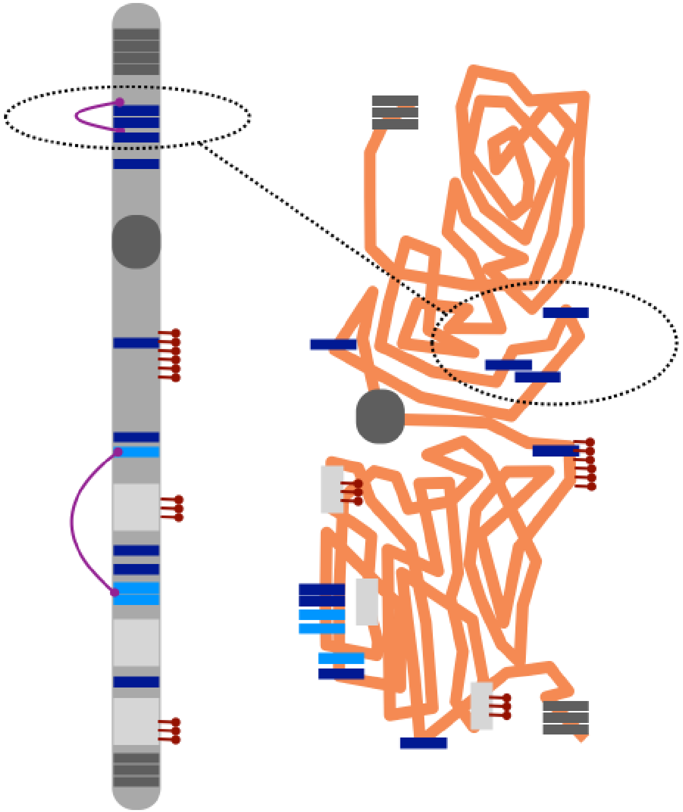
- ⊂ Chromatin conformation
- ⇶ Epigenetic marks

Genes

- ▬ Protein coding genes.
- ▬ Genes producing ncRNA

Non-functional ?

- ■ Repetitive elements



Types of ”annotation”

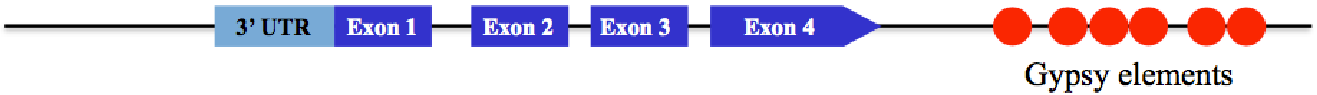
Structural annotation consists of the identification of genomic elements.

- ORFs and their localization
- gene structure
- coding regions
- location of regulatory motifs
- repeats

Functional annotation consists of attaching biological information to genomic elements.

- biochemical function
- biological function
- involved regulation and interactions
- expression

Structural annotation



Functional annotation

kinase

Annotation Strategies

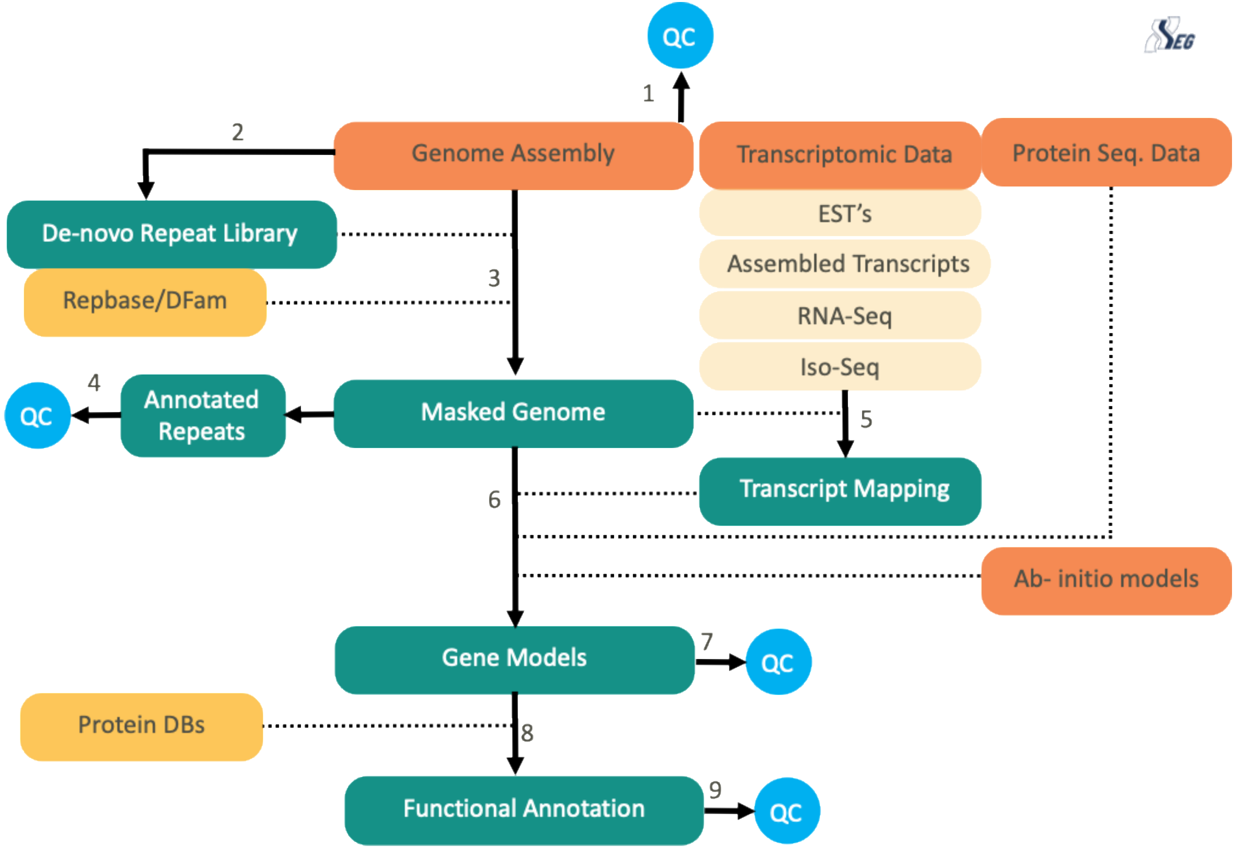
Annotation types

- **Automatic annotation** uses pattern recognition algorithms like Markov Chain models.
- Quality depends on training data; effective for repetitive elements but limited for complex unique genes.
- **Manual annotation** involves human-supervised inspection, producing high-quality functional annotations.
- Manual annotation is feasible mainly in model organisms such as *Saccharomyces cerevisiae* and *Arabidopsis thaliana*.

Identification Genomics Elements

- Sequence homology methods include RNA-Seq data and known transposon sequences.
- **Pattern-based recognition** detects motifs such as ATG and GAx sequences.
- Experimental data sources include **HiC** contact maps and methylation pattern analyses.
- **Combining** multiple approaches enhances annotation accuracy and reliability.

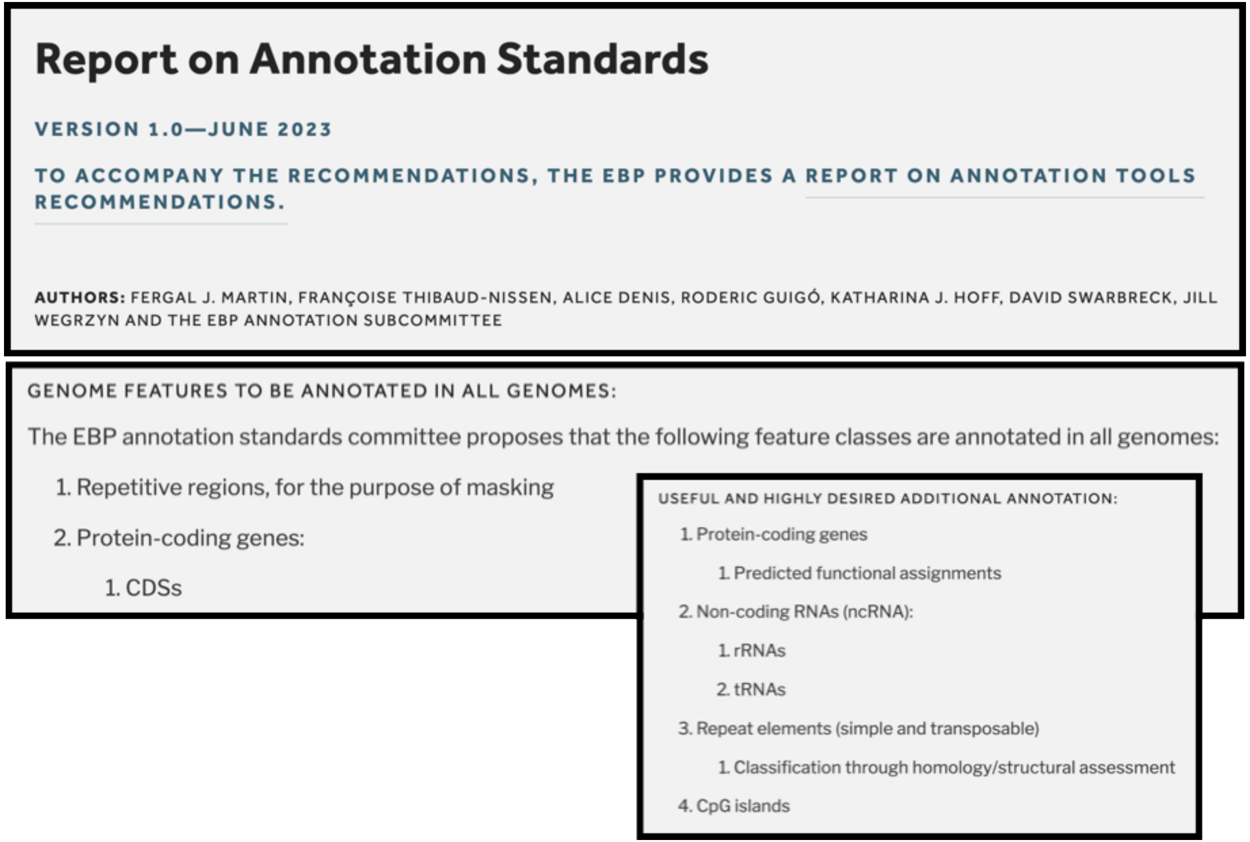
Steps in gene annotation



Annotation Standards

EARTH BIOGENOME PROJECT

https://www.earthbiogenome.org/



Report on Annotation Standards

VERSION 1.0—JUNE 2023

TO ACCOMPANY THE RECOMMENDATIONS, THE EBP PROVIDES A REPORT ON ANNOTATION TOOLS RECOMMENDATIONS.

AUTHORS: FERGAL J. MARTIN, FRANÇOISE THIBAUD-NISSEN, ALICE DENIS, RODERIC GUIGÓ, KATHARINA J. HOFF, DAVID SWARBRECK, JILL WEGRZYN AND THE EBP ANNOTATION SUBCOMMITTEE

GENOME FEATURES TO BE ANNOTATED IN ALL GENOMES:

The EBP annotation standards committee proposes that the following feature classes are annotated in all genomes:

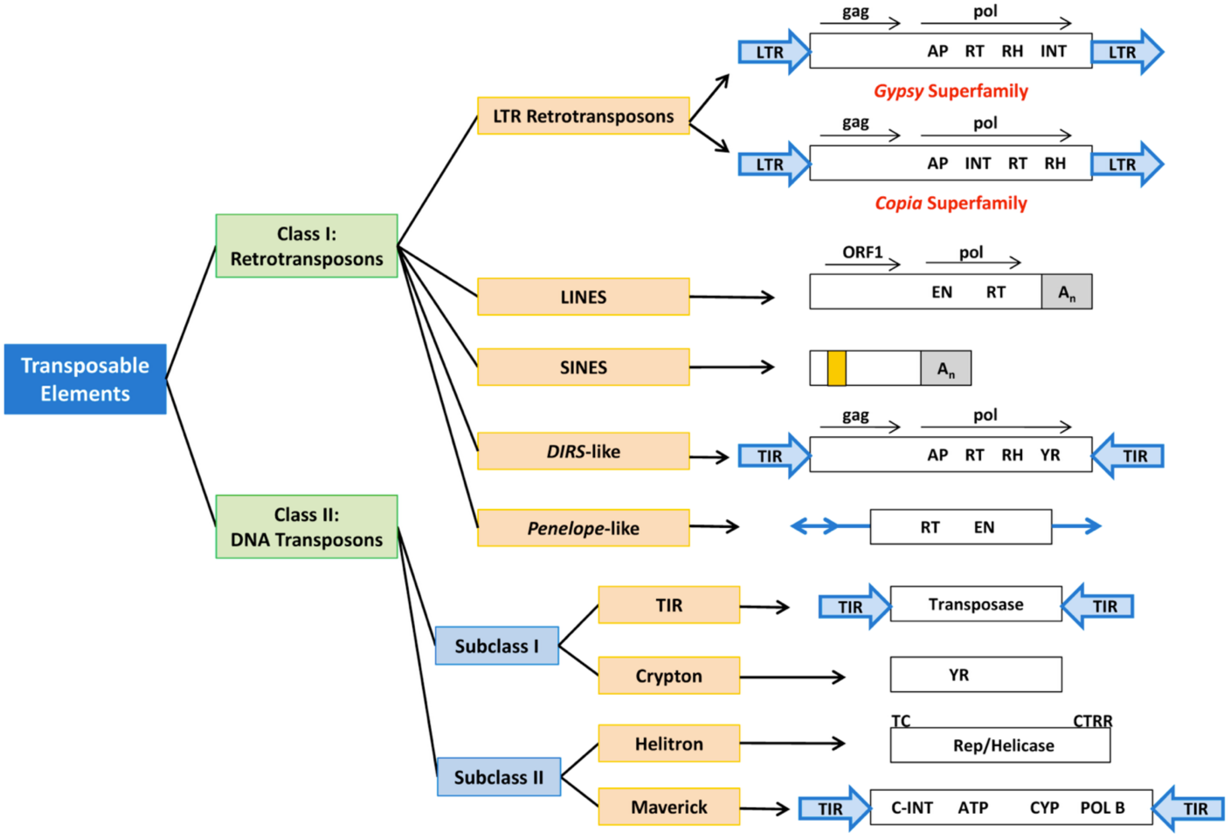
1. Repetitive regions, for the purpose of masking
2. Protein-coding genes:
   1. CDSs

USEFUL AND HIGHLY DESIRED ADDITIONAL ANNOTATION:

1. Protein-coding genes
   1. Predicted functional assignments
2. Non-coding RNAs (ncRNA):
   1. rRNAs
   2. tRNAs
3. Repeat elements (simple and transposable)
   1. Classification through homology/structural assessment
4. CpG islands

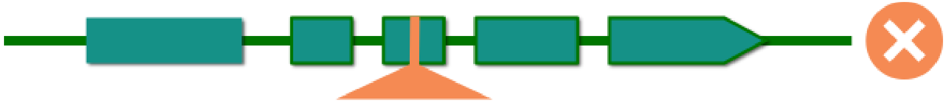
Annotation of Repetitive Elements

Repeated sequences (repetitive elements, or repeats) are patterns of nucleic acids (DNA or RNA) **that occur in multiple copies throughout the genome**. The functions and descriptions of these sequences are currently being characterized by scientists. Repetitive DNA was first detected because of its rapid reassociation kinetics.

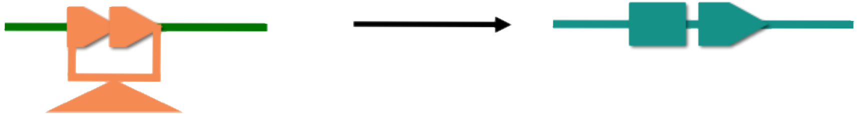


Relevance of Repetitive Elements

They have effects on gene function, from altering its expression to disrupt its function and convert it in a pseudogene



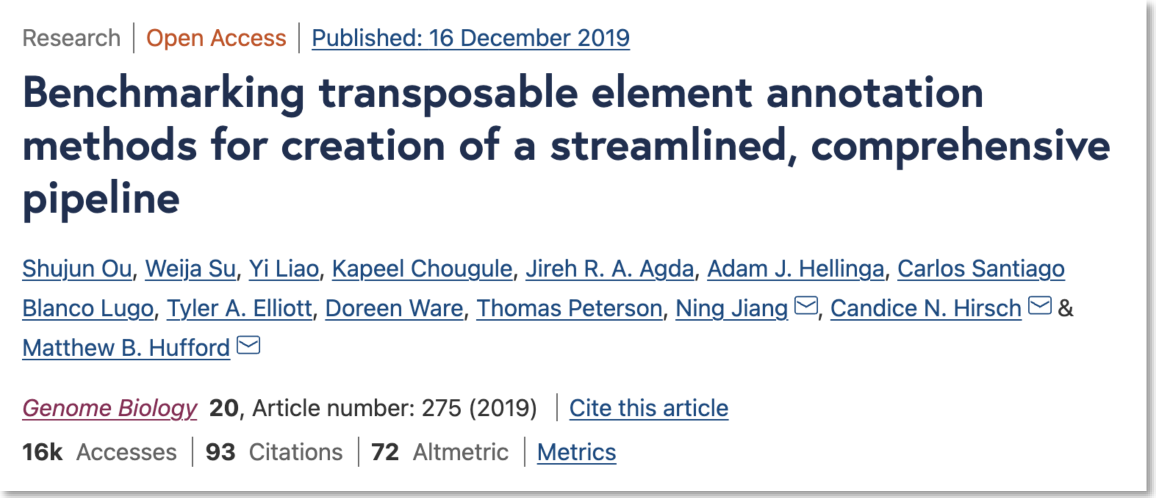
They can be domesticated being a possible source of new genes



They are an important source of the genome dynamics, from recombination to generation of new genomic elements.

Software to annotate RE

PROGRAM	TYPE	APPROACH	CITATION
RepeatMasker	Library based.	Search by homology	Smit et al. 1996
PLOTREP (Censor)	Library based.	Search by homology	Toth et al. 2006
LTR_STRUCT	Library based.	Search for LTR Transposons	McCarthy and McDonald 2003
Greedier	Library based.	Search by homology. Nested elements	Li et al. 2008
RTAnalyzer	Signature based	LINEs, Alus and retrogenes using Blast	Lucier et al. 2007
FINDMITE	Signature based		
HelitronFinder	Signature based		
LTR_Retriever	Ab-initio		
RECON	Ab-initio		
PILER	Ab-initio		
RepeatScout	Ab-initio		
RepeatFinder/REPuter	Ab-initio		
RepeatRunner	Pipeline		
RepeatModeler2	Pipeline		
EDTA	Pipeline		
REPET	Pipeline	Combination of many tools	Hoede et al. 2014



Research | Open Access | Published: 16 December 2019

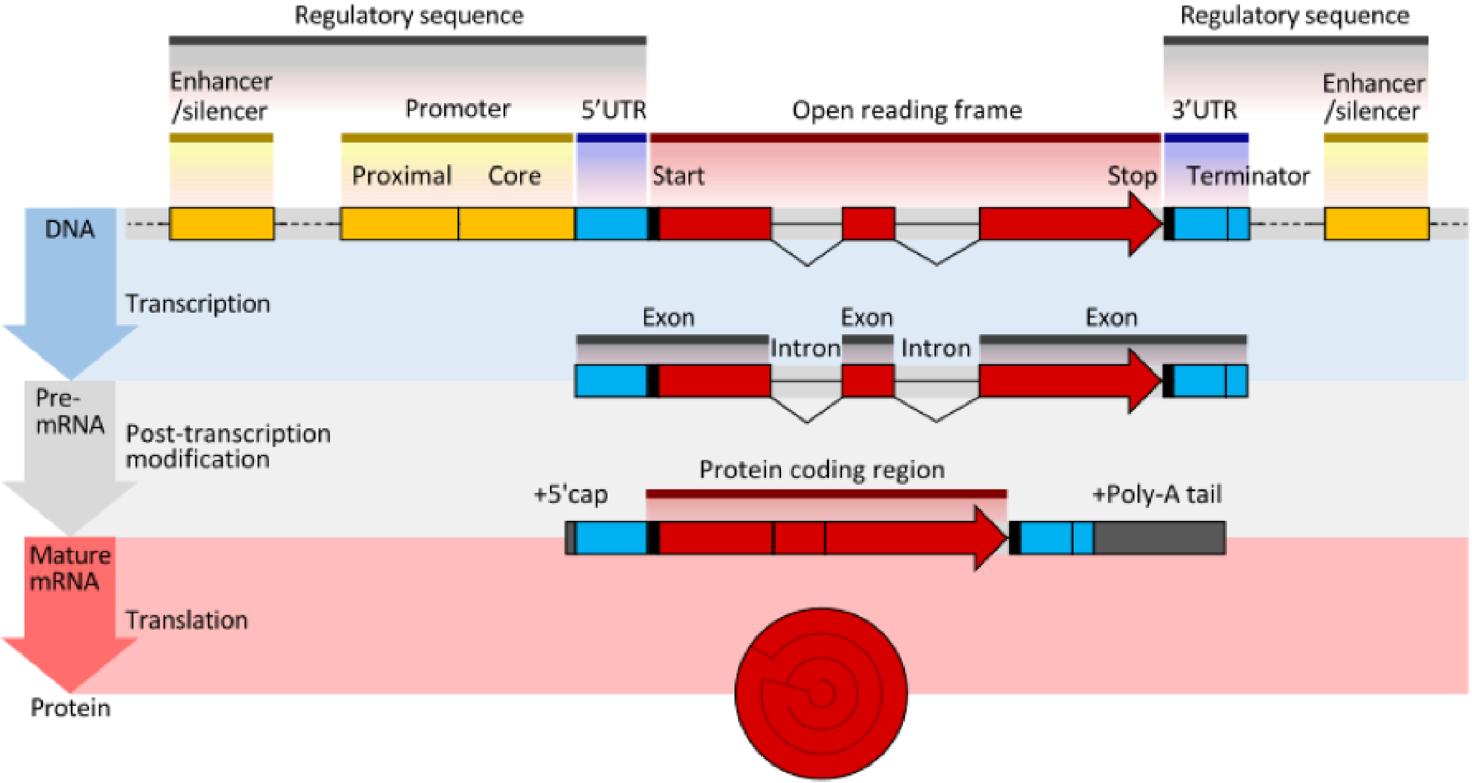
Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline

Shujun Ou, Weija Su, Yi Liao, Kapeel Chougule, Jireh R. A. Agda, Adam J. Hellinga, Carlos Santiago Blanco Lugo, Tyler A. Elliott, Doreen Ware, Thomas Peterson, Ning Jiang, Candice N. Hirsch & Matthew B. Hufford

Genome Biology **20**, Article number: 275 (2019) | Cite this article

16k Accesses | **93** Citations | **72** Altmetric | Metrics

Gene Annotation



Gene Annotation Strategies

Ab-initio

- Rely in mathematical models to determine intron-exon structure.
- Do not external evidence (e.g. ESTs).
- Do not report untranslated regions (UTRs).
- Accuracy intron-exon structure < 60%.

Evidence-based

- ESTs, RNA-seq and know protein data need to be aligned.
- Good accuracy.
- Poorer sensitivity.
- Computationally intensive.

Evidence-driven

- Does first *ab initio*, then refine with experimental data
- Combine the best of the both worlds
- Improves sensitivity
- Computationally intensive.

Keys for a successful annotation

Good assembly

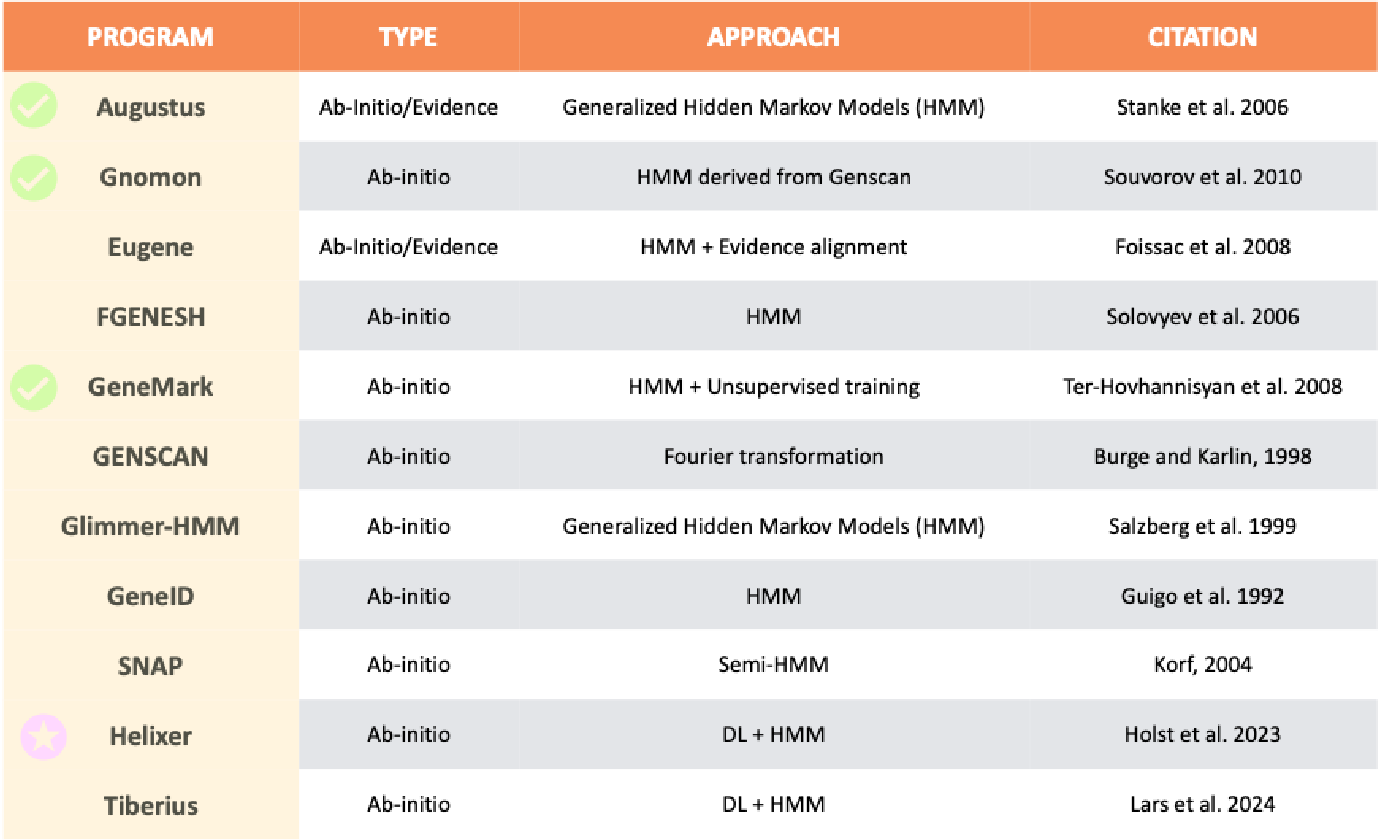
Good TE Library

Much transcript / protein data

Good *ab initio* models

Quality data from other species

Ab initio gene prediction algorithms



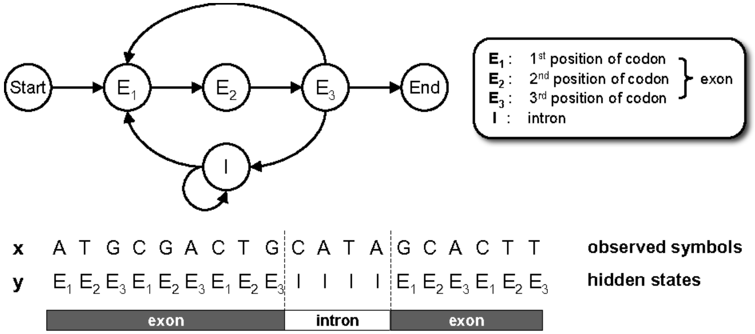
PROGRAM	TYPE	APPROACH	CITATION
✓ Augustus	Ab-Initio/Evidence	Generalized Hidden Markov Models (HMM)	Stanke et al. 2006
✓ Gnomon	Ab-initio	HMM derived from Genscan	Souvorov et al. 2010
Eugene	Ab-Initio/Evidence	HMM + Evidence alignment	Foissac et al. 2008
FGENESH	Ab-initio	HMM	Solovyev et al. 2006
✓ GeneMark	Ab-initio	HMM + Unsupervised training	Ter-Hovhannisyan et al. 2008
GENSCAN	Ab-initio	Fourier transformation	Burge and Karlin, 1998
Glimmer-HMM	Ab-initio	Generalized Hidden Markov Models (HMM)	Salzberg et al. 1999
GeneID	Ab-initio	HMM	Guigo et al. 1992
SNAP	Ab-initio	Semi-HMM	Korf, 2004
★ Helixer	Ab-initio	DL + HMM	Holst et al. 2023
Tiberius	Ab-initio	DL + HMM	Lars et al. 2024

Ab initio gene prediction algorithms

There are two types of algorithms applied to gene structure identification:

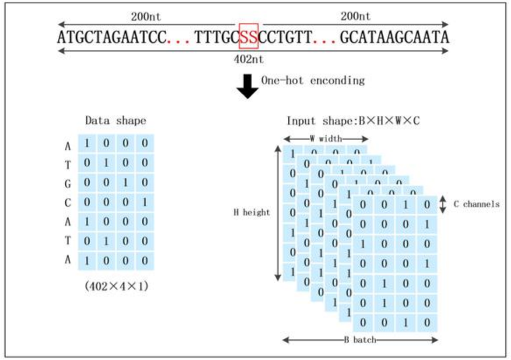
- Hidden Markov Models (HMM).
  - Examples: Augustus, Gnomon…

Usually needs to be trained by SPECIES
(They do not generalise well)


Start
E_1
E_2
E_3
End
I
E_1 : 1st position of codon
E_2 : 2nd position of codon
E_3 : 3rd position of codon
exon
I : intron
x A T G C G A C T G C A T A G C A C T T observed symbols
y E_1 E_2 E_3 E_1 E_2 E_3 E_1 E_2 E_3 I I I I E_1 E_2 E_3 E_1 E_2 E_3 hidden states
exon
intron
exon

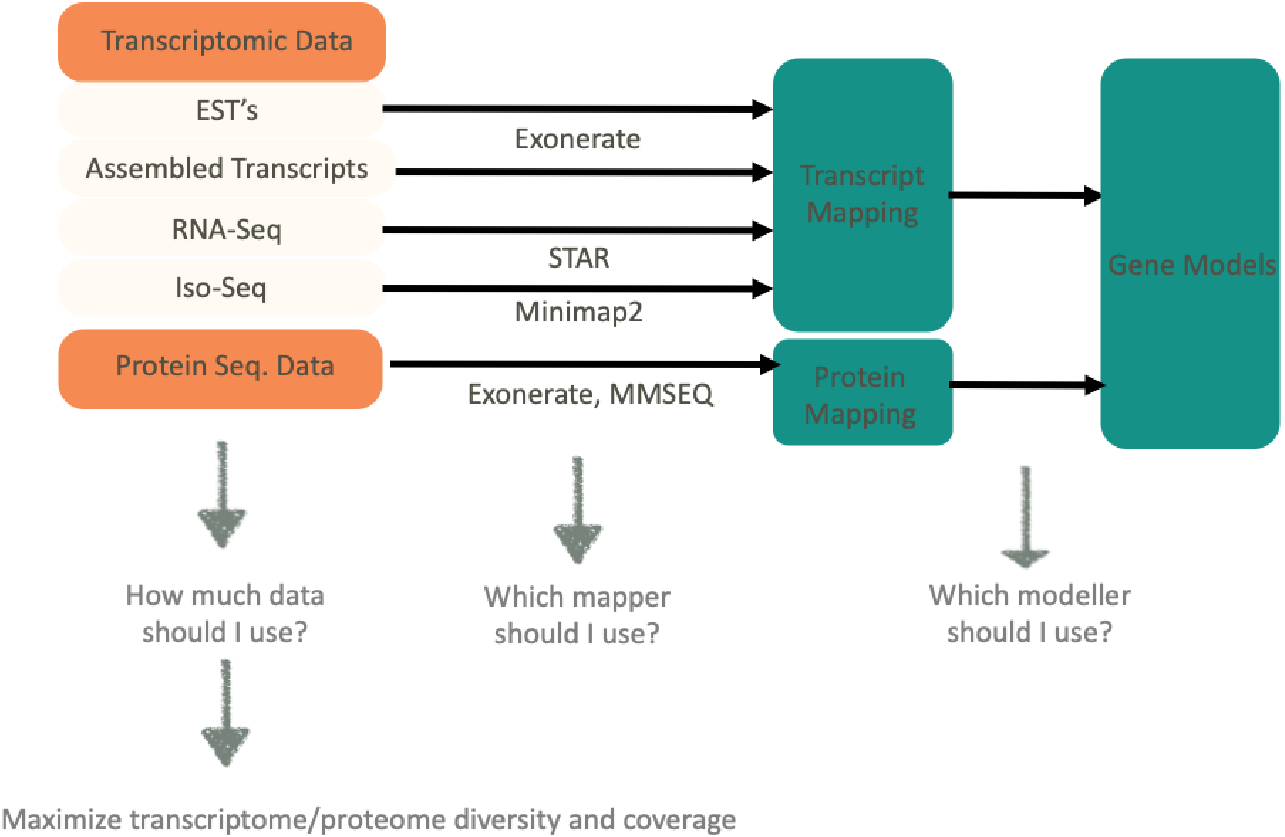
- Neural Networks and Deep Learning Models (DL).
  - Example: Helixer, Tiberius…

They can be trained by LINEAGE
They need GPUs to be efficient



A	1	0	0	0
T	0	1	0	0
G	0	0	1	0
C	0	0	0	1
A	1	0	0	0
T	0	1	0	0
A	1	0	0	0

Evidence-based gene prediction



Transcriptomic Data

EST's

Assembled Transcripts

RNA-Seq

Iso-Seq

Exonerate

STAR

Minimap2

Protein Seq. Data

Exonerate, MMSEQ

Transcript Mapping

Protein Mapping

Gene Models

How much data should I use?

Which mapper should I use?

Which modeller should I use?

Maximize transcriptome/proteome diversity and coverage

Evidence based gene prediction algorithms

PROGRAM	TYPE	APPROACH	CITATION
Exonerate	Transcripts or Proteins Evidence (EvT, EvP)	Sequence alignment (used by Maker)	Slater and Birney, 2005
PASA	Transcripts Evidence (EvT)	Transcript model assembly	Haas et al. 2003
✓ Tophat/Cufflinks	SR Transcripts Evidence (EvT)	Based on RNA-Seq alignments	Trapnell et al. 2012
GeneWise	Protein Sequence Evidence (EvP)	Sequence alignment (obsolete)	Birney et al. 2004
GenomeScan	Protein Sequence Evidence (EvP)	Sequence alignment (obsolete)	Yeh et al. 2001
✓ TransDeCoder	Protein Sequence Evidence (EvP)	Based on the longest ORF + Sequence homology hits (BLAST/HMMSCAN)	NA
✓ T2D	Protein Sequence Evidence (EvP)	New version of TransDeCoder	Mao et al. 2025
✓ GeMoMa	Protein Sequence Evidence (EvP)	Protein alignment (+ opt. transcriptomic data)	Keilwagen et al. 2019

Evaluation of annotation methods

Completeness

Do I have capture the whole gene space?

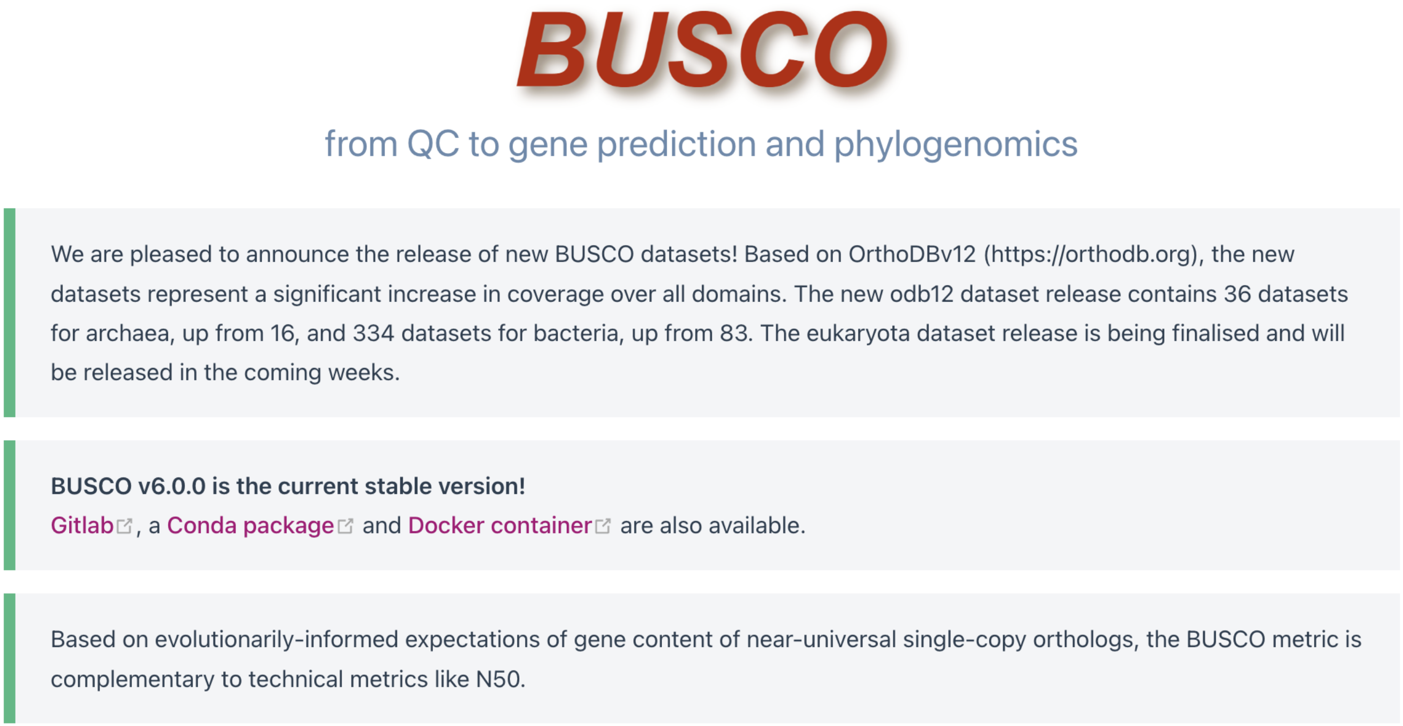
Precision

Are my gene models correct on intron-exon, CDS and UTRs, and nucleotides?

Contaminations

Are my gene models real genes or do I have pseudogenes, TE… identified as genes?

BUSCO genes



BUSCO

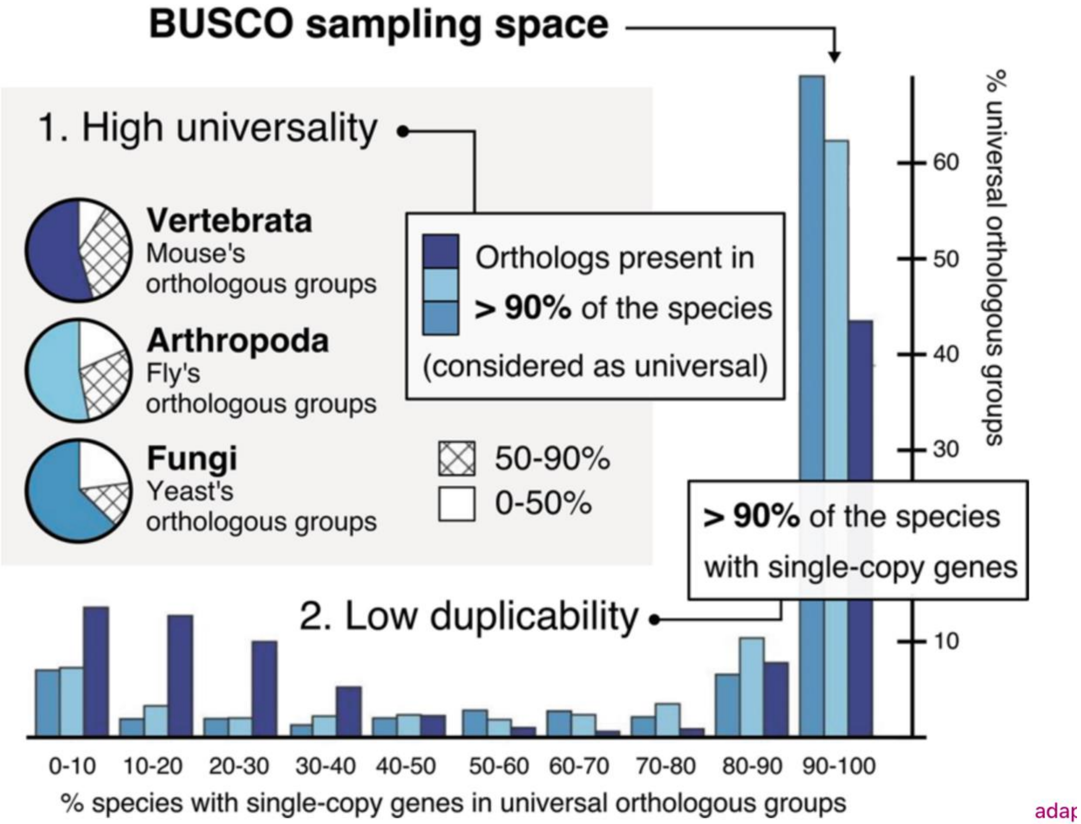
from QC to gene prediction and phylogenomics

> We are pleased to announce the release of new BUSCO datasets! Based on OrthoDBv12 (https://orthodb.org), the new datasets represent a significant increase in coverage over all domains. The new odb12 dataset release contains 36 datasets for archaea, up from 16, and 334 datasets for bacteria, up from 83. The eukaryota dataset release is being finalised and will be released in the coming weeks.

> **BUSCO v6.0.0 is the current stable version!**
> Gitlab, a Conda package and Docker container are also available.

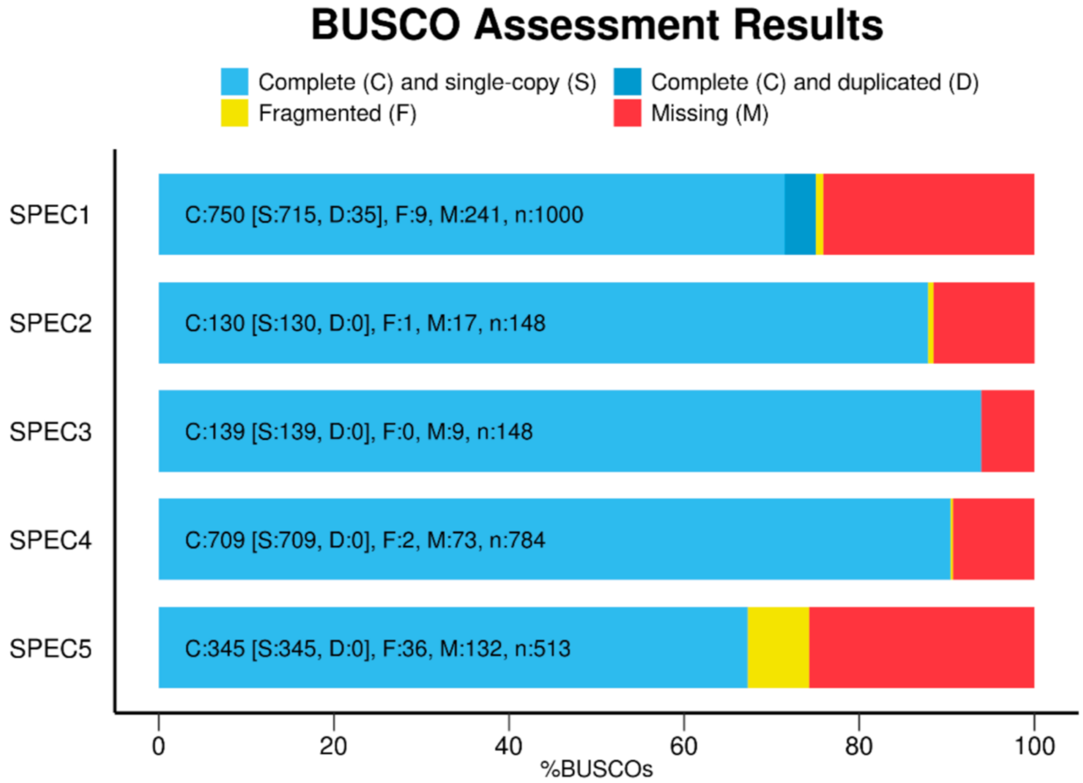
> Based on evolutionarily-informed expectations of gene content of near-universal single-copy orthologs, the BUSCO metric is complementary to technical metrics like N50.

BUSCO genes

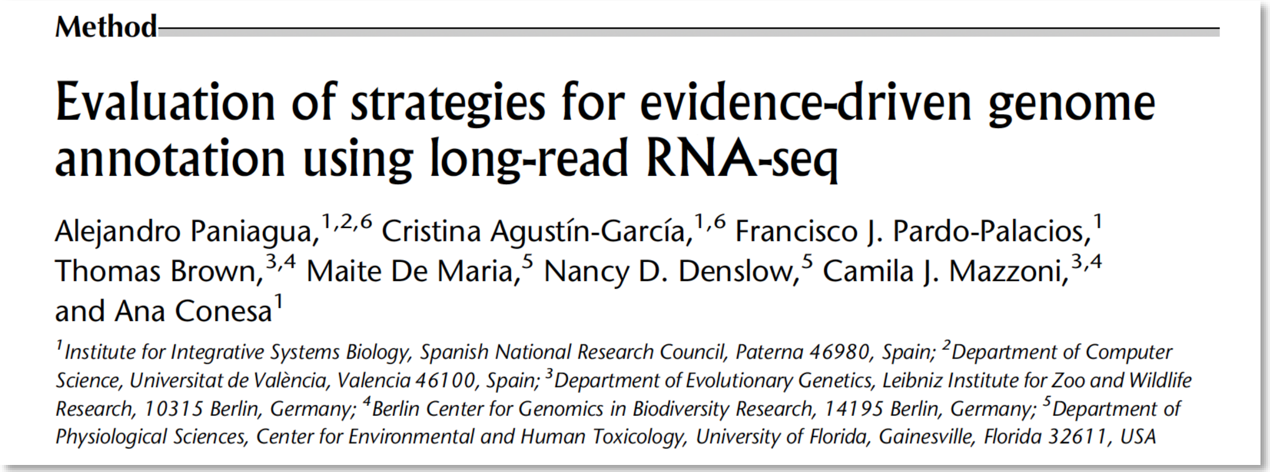


adap

BUSCO assessment



Evidence-driven Strategies



Method

Evaluation of strategies for evidence-driven genome annotation using long-read RNA-seq

Alejandro Paniagua,[1,2,6] Cristina Agustín-García,[1,6] Francisco J. Pardo-Palacios,[1] Thomas Brown,[3,4] Maite De Maria,[5] Nancy D. Denslow,[5] Camila J. Mazzoni,[3,4] and Ana Conesa[1]

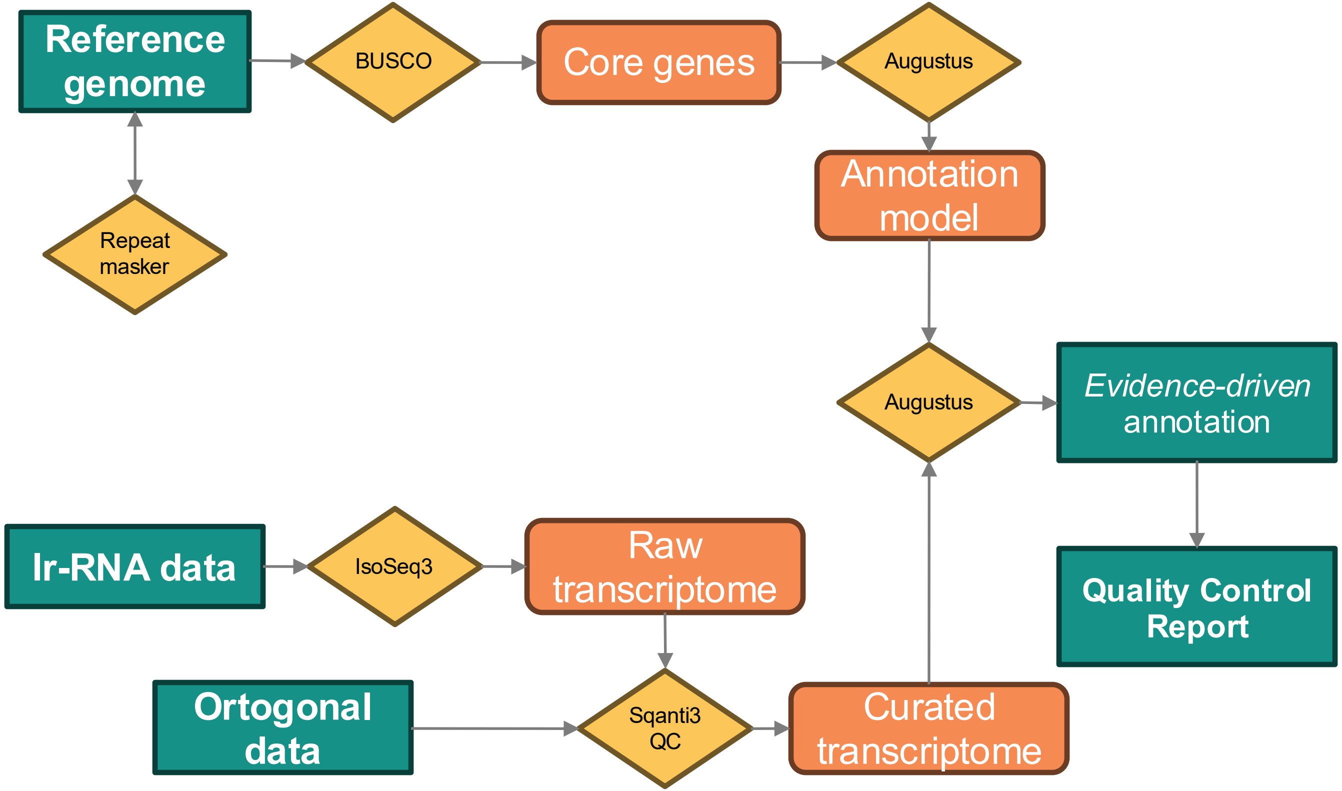
[1] *Institute for Integrative Systems Biology, Spanish National Research Council, Paterna 46980, Spain;* [2] *Department of Computer Science, Universitat de València, Valencia 46100, Spain;* [3] *Department of Evolutionary Genetics, Leibniz Institute for Zoo and Wildlife Research, 10315 Berlin, Germany;* [4] *Berlin Center for Genomics in Biodiversity Research, 14195 Berlin, Germany;* [5] *Department of Physiological Sciences, Center for Environmental and Human Toxicology, University of Florida, Gainesville, Florida 32611, USA*

Evidence-driven

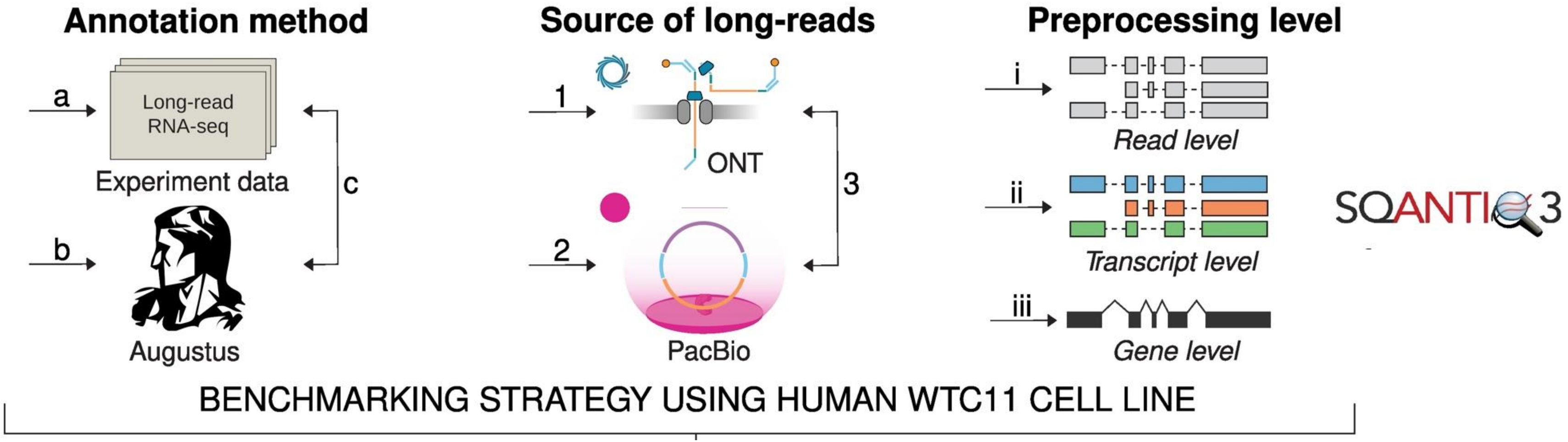
- Does first *ab initio*, then refine with experimental data
- Combine the best of the both worlds
- Improves sensitivity
- Computationally intensive.

Paniagua *et al.* **Genome Research**, *2025*

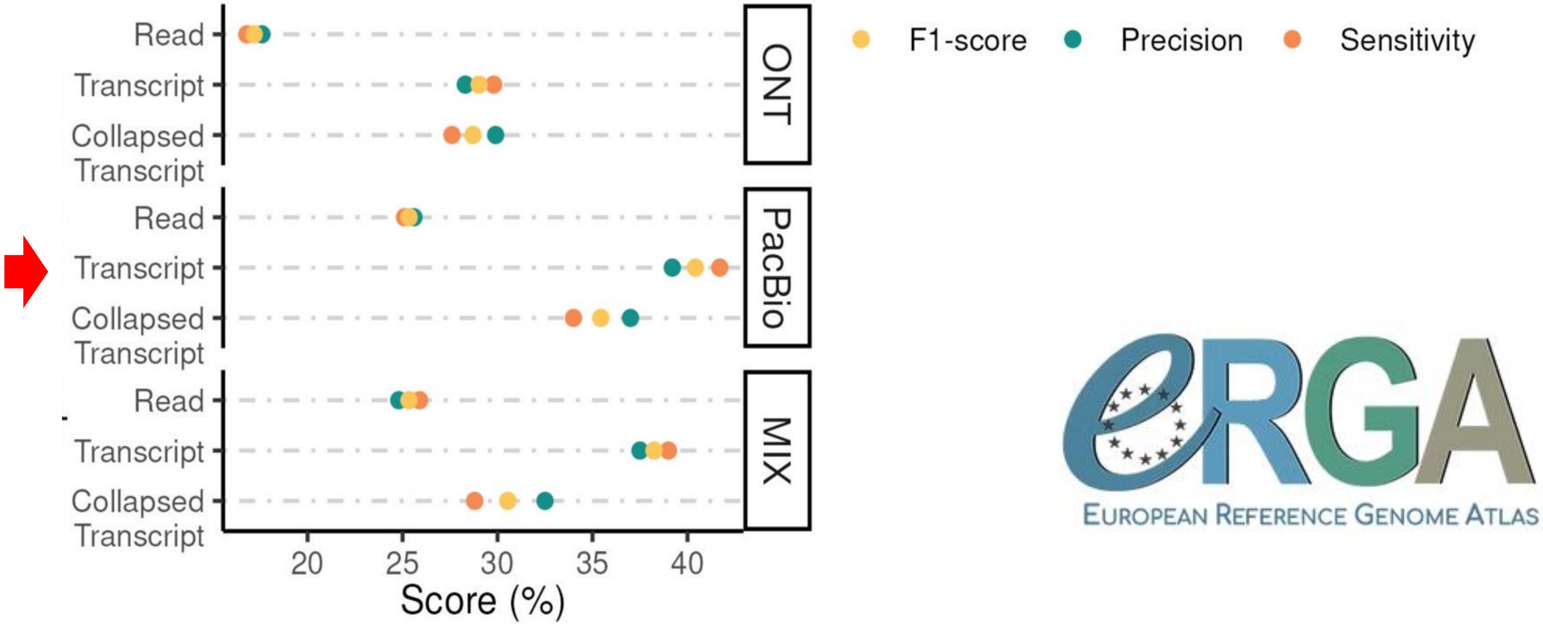
Genome annotation and lrRNA-seq



Genome annotation supported by lrRNA-seq



Curated and reconstructed transcripts outperforms



EUROPEAN REFERENCE GENOME ATLAS

Amount of data needed

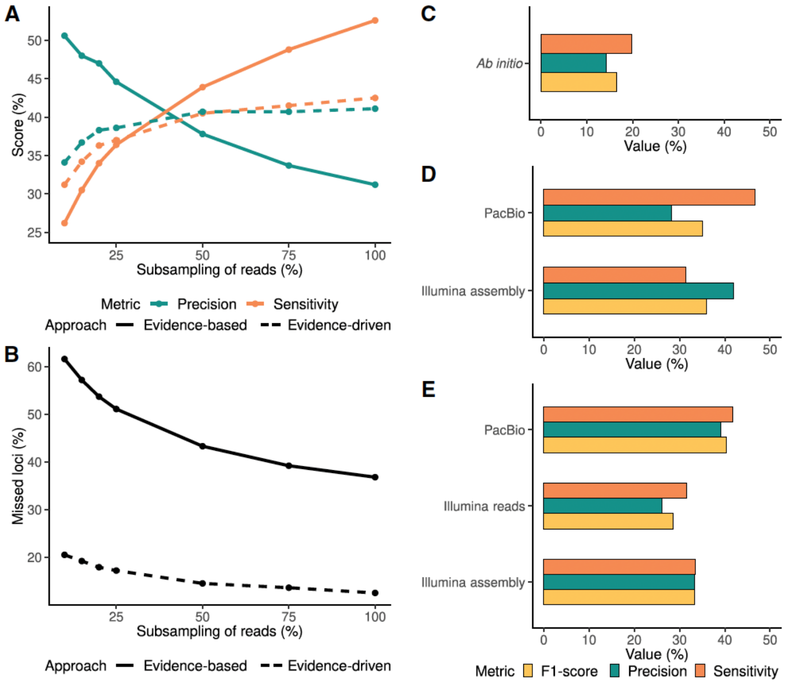
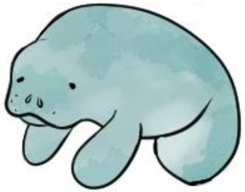
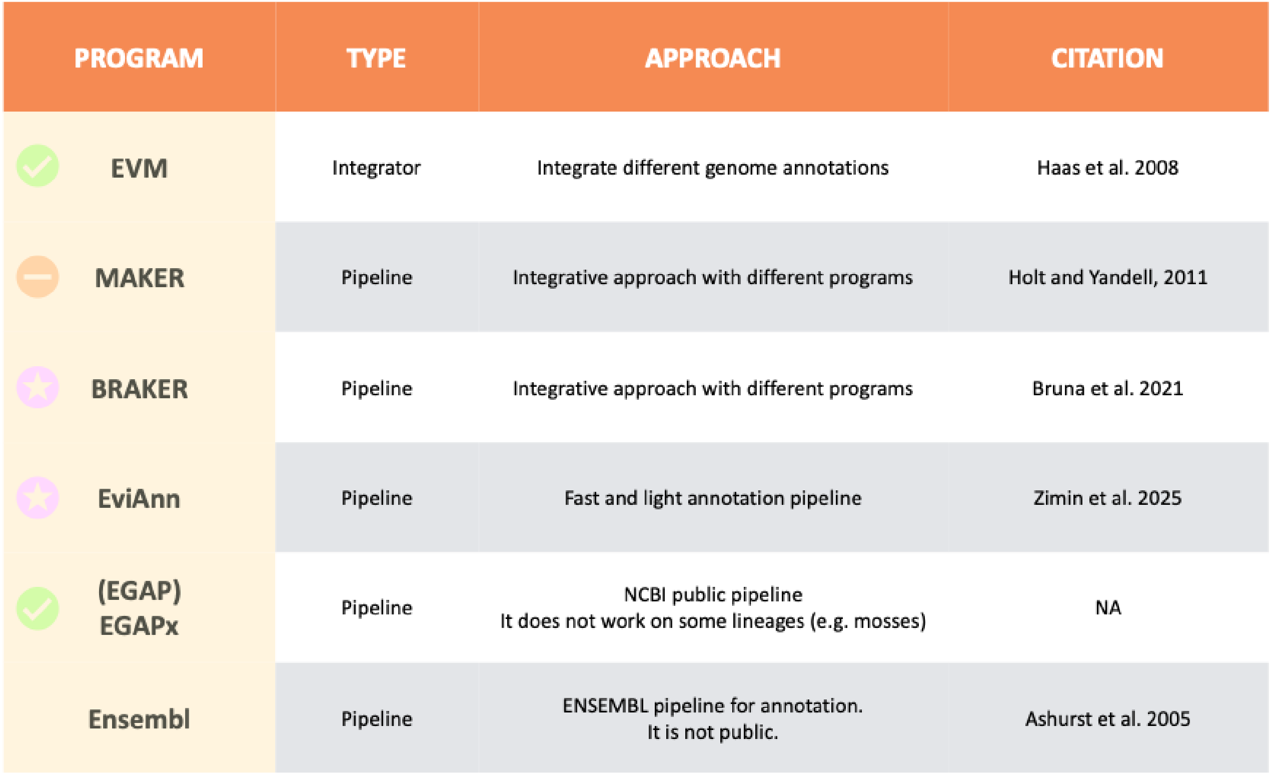


Figure 3. Performance analysis of gene prediction as a function of the number of reads. Sensitivity, precision (*A*), and number of missed loci (*B*) were obtained with different sample sizes of WTC11 cell line PacBio FLNC reads using evidence-based and evidence-driven approaches. Performance of the different genome annotation approaches with Illumina short-read and PacBio long-read technologies at the gene level. (C) Ab initio predictions. (*D*) Evidence-based models using PacBio and Illumina-assembled transcriptomes. (*E*) Evidence-driven approach with PacBio, Illumina reads or Illumina-assembled transcriptomes as the source of evidence for the prediction step.

Application to the manatee genome annotation

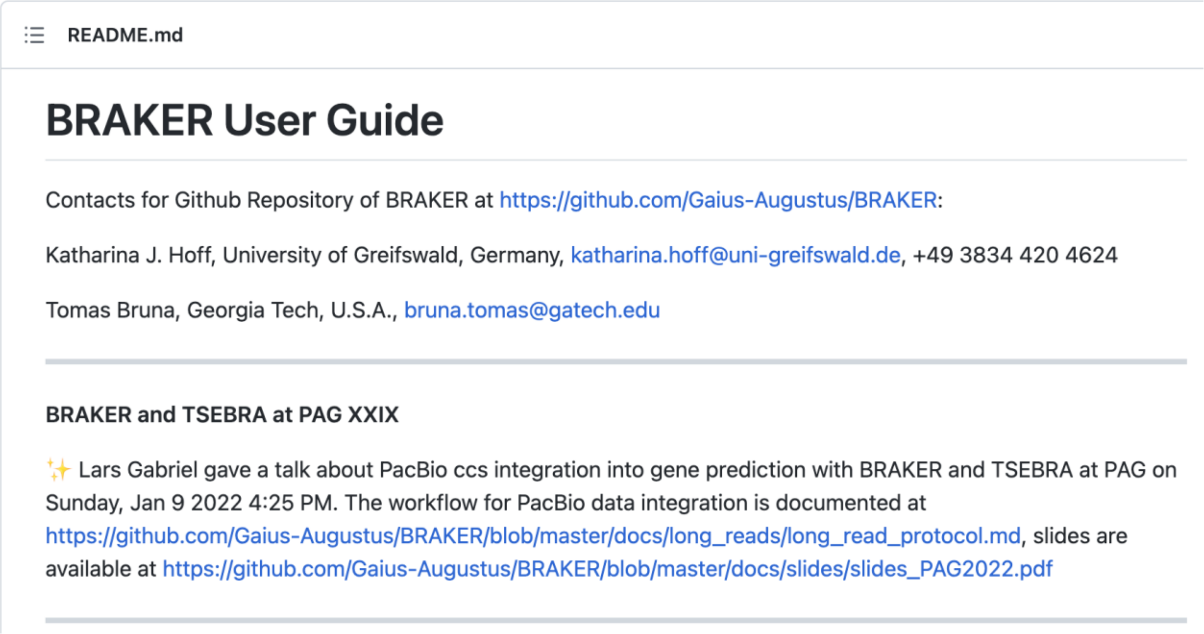


Genome annotation pipelines



PROGRAM	TYPE	APPROACH	CITATION
✓ EVM	Integrator	Integrate different genome annotations	Haas et al. 2008
⊖ MAKER	Pipeline	Integrative approach with different programs	Holt and Yandell, 2011
☆ BRAKER	Pipeline	Integrative approach with different programs	Bruna et al. 2021
☆ EviAnn	Pipeline	Fast and light annotation pipeline	Zimin et al. 2025
✓ (EGAP) EGAPx	Pipeline	NCBI public pipeline It does not work on some lineages (e.g. mosses)	NA
Ensembl	Pipeline	ENSEMBL pipeline for annotation. It is not public.	Ashurst et al. 2005

Some recommended tools



README.md

BRAKER User Guide

Contacts for Github Repository of BRAKER at https://github.com/Gaius-Augustus/BRAKER:

Katharina J. Hoff, University of Greifswald, Germany, katharina.hoff@uni-greifswald.de, +49 3834 420 4624

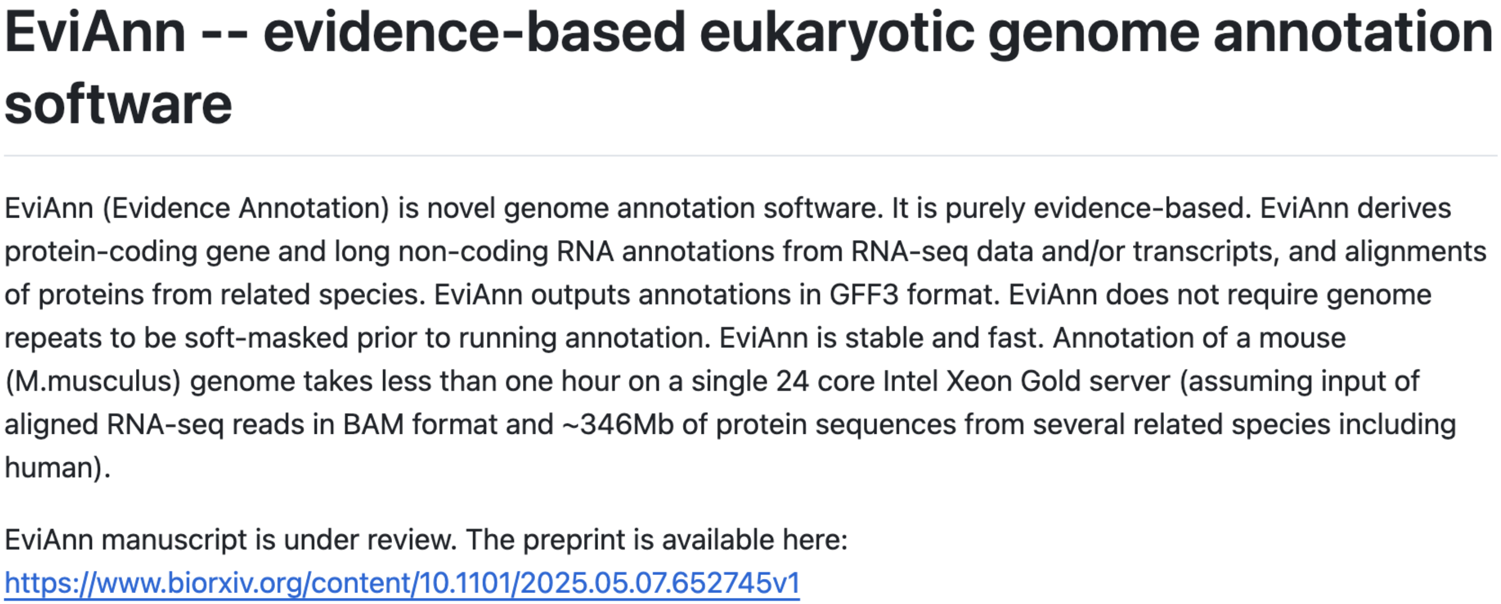
Tomas Bruna, Georgia Tech, U.S.A., bruna.tomas@gatech.edu

BRAKER and TSEBRA at PAG XXIX

✨ Lars Gabriel gave a talk about PacBio ccs integration into gene prediction with BRAKER and TSEBRA at PAG on Sunday, Jan 9 2022 4:25 PM. The workflow for PacBio data integration is documented at https://github.com/Gaius-Augustus/BRAKER/blob/master/docs/long_reads/long_read_protocol.md, slides are available at https://github.com/Gaius-Augustus/BRAKER/blob/master/docs/slides/slides_PAG2022.pdf

---

Some recommended tools



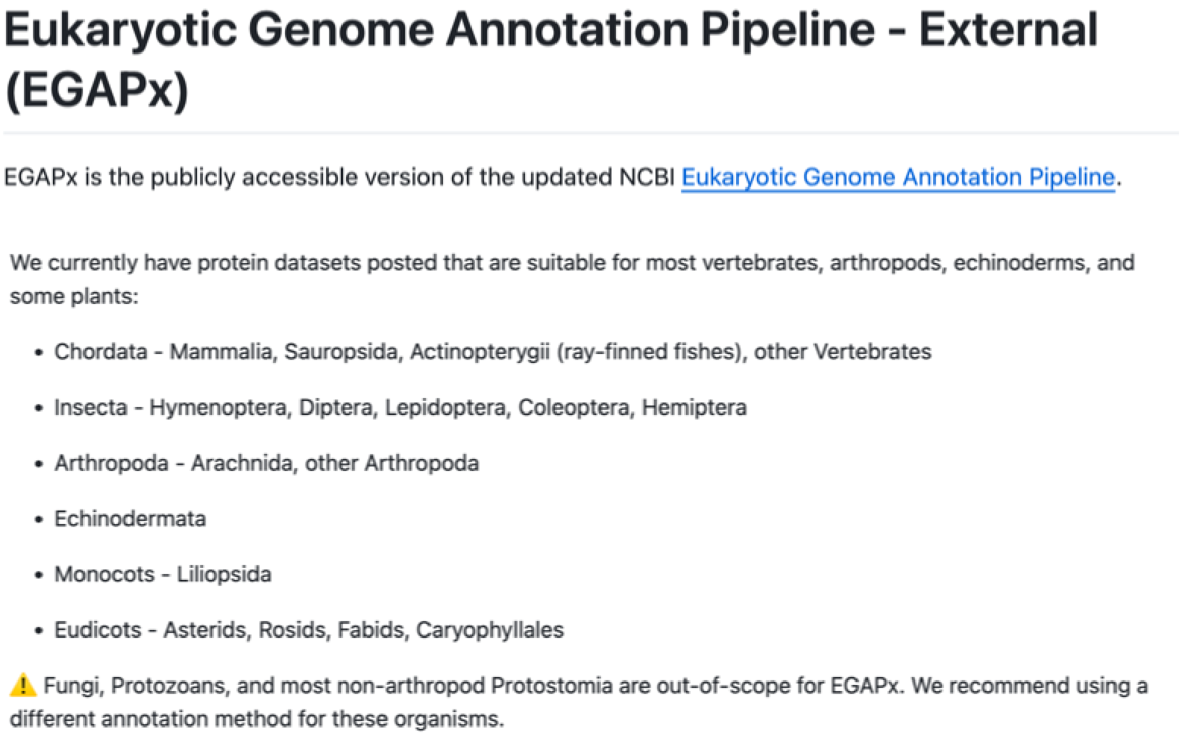
EviAnn -- evidence-based eukaryotic genome annotation software

EviAnn (Evidence Annotation) is novel genome annotation software. It is purely evidence-based. EviAnn derives protein-coding gene and long non-coding RNA annotations from RNA-seq data and/or transcripts, and alignments of proteins from related species. EviAnn outputs annotations in GFF3 format. EviAnn does not require genome repeats to be soft-masked prior to running annotation. EviAnn is stable and fast. Annotation of a mouse (M.musculus) genome takes less than one hour on a single 24 core Intel Xeon Gold server (assuming input of aligned RNA-seq reads in BAM format and ~346Mb of protein sequences from several related species including human).

EviAnn manuscript is under review. The preprint is available here: https://www.biorxiv.org/content/10.1101/2025.05.07.652745v1

Some recommended tools

https://github.com/ncbi/egapx



Eukaryotic Genome Annotation Pipeline - External (EGAPx)

EGAPx is the publicly accessible version of the updated NCBI Eukaryotic Genome Annotation Pipeline.

We currently have protein datasets posted that are suitable for most vertebrates, arthropods, echinoderms, and some plants:

- Chordata - Mammalia, Sauropsida, Actinopterygii (ray-finned fishes), other Vertebrates
- Insecta - Hymenoptera, Diptera, Lepidoptera, Coleoptera, Hemiptera
- Arthropoda - Arachnida, other Arthropoda
- Echinodermata
- Monocots - Liliopsida
- Eudicots - Asterids, Rosids, Fabids, Caryophyllales

⚠ Fungi, Protozoans, and most non-arthropod Protostomia are out-of-scope for EGAPx. We recommend using a different annotation method for these organisms.

Check your knowledge

- What is the difference between evidence-driven and ab inition gene predicton methods
- Is gene annotation all genome annotation?
- What is BUSCO?
- Indicate 3 key element of success for geneome annotation
- What can you do to annotate a genome if you do not have much experimental evidence?
- What does fragmented BUSCO mean?