Tools for Annotation

Analysis	Tool
Ab initio gene prediction	Augustus Helixer (deep learning)
Repeat Discovery/ Masking/ Classification	RepeatModeler2 RepeatMasker Repeat Detector Dust TRF - Tandem Repeat Finder
Short-read transcriptomic data alignment (splice-aware)	STAR Splign
Transcript assembly (reference based)	Scallop Stringtie Cufflinks
Transcript assembly (de novo)	Trinity
Long-read transcriptomic data alignment	MiniMap2

Analysis	Tool
Validating protein-coding transcripts (against a protein database)	BLAST Diamond
Validating protein-coding transcripts (ML approaches)	CPC2 RNASamba RNAmining
Protein-to-genome Alignment	GenBLAST MiniProt Pro-splign
Whole genome alignment	LastZ Cactus
Map annotation between genomes	LiftOff
Genome annotation editor	Apollo3
Assess proteome completeness	BUSCO OMArk

Tools for Annotation

Pipeline	Description
BRAKER3	*BRAKER3 is the latest pipeline in the BRAKER suite. It enables the usage of RNA-seq and protein data in a fully automated pipeline to train and predict highly reliable genes with GeneMark-ETP and AUGUSTUS. The result of the pipeline is the combined gene set of both gene prediction tools, which only contains genes with very high support from extrinsic evidence.*
GALBA	*A fully automated gene prediction pipeline that trains AUGUSTUS for a novel species and subsequently predicts genes with AUGUSTUS in the genome of that species. GALBA uses the protein sequences of one closely related species to generate a training gene set for AUGUSTUS with either miniprot, or GenomeThreader. After training, GALBA uses the evidence from protein to genome alignment during gene prediction.*
MAKER	*MAKER can be used for de novo annotation of newly sequenced genomes, for updating existing annotations to reflect new evidence, or just to combine annotations, evidence, and quality control statistics for use with other GMOD programs*
nf-core/genomeannotator (in development)	*nf-core/genomeannotator combines a number of established tools for the assembly, alignment and subsequent integration of so-called evidences into consensus gene builds. The product of nf-core/genomeannotator are various tracks in GFF format, including gene models, but also various alignments. Output from nf-core/genomeannotator is largely compatible with GMOD.*
Ensembl-anno (very much still in development)	At Ensembl, we are working on an Annotation toolkit, ensembl-anno. Currently, it exists as a large Python script that requires paths to software and input data to run several analyses for annotating a genome. It is still somewhat tied to Ensembl infrastructure, locally installed software and dependencies on the Ensembl Perl API. The goal of this project is to turn the individual analyses into modules that can be called from NextFlow pipelines and run on multiple genomes in parallel. The entire toolkit will be fully deployable and work with containerised software, so that it can be run by anyone anywhere. Ensembl-anno… coming 2024(?)... watch this space!