The NCBI Eukaryotic Genome Annotation Process

Enhancing the value of assembled genomes through annotation using a standardized pipeline

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

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

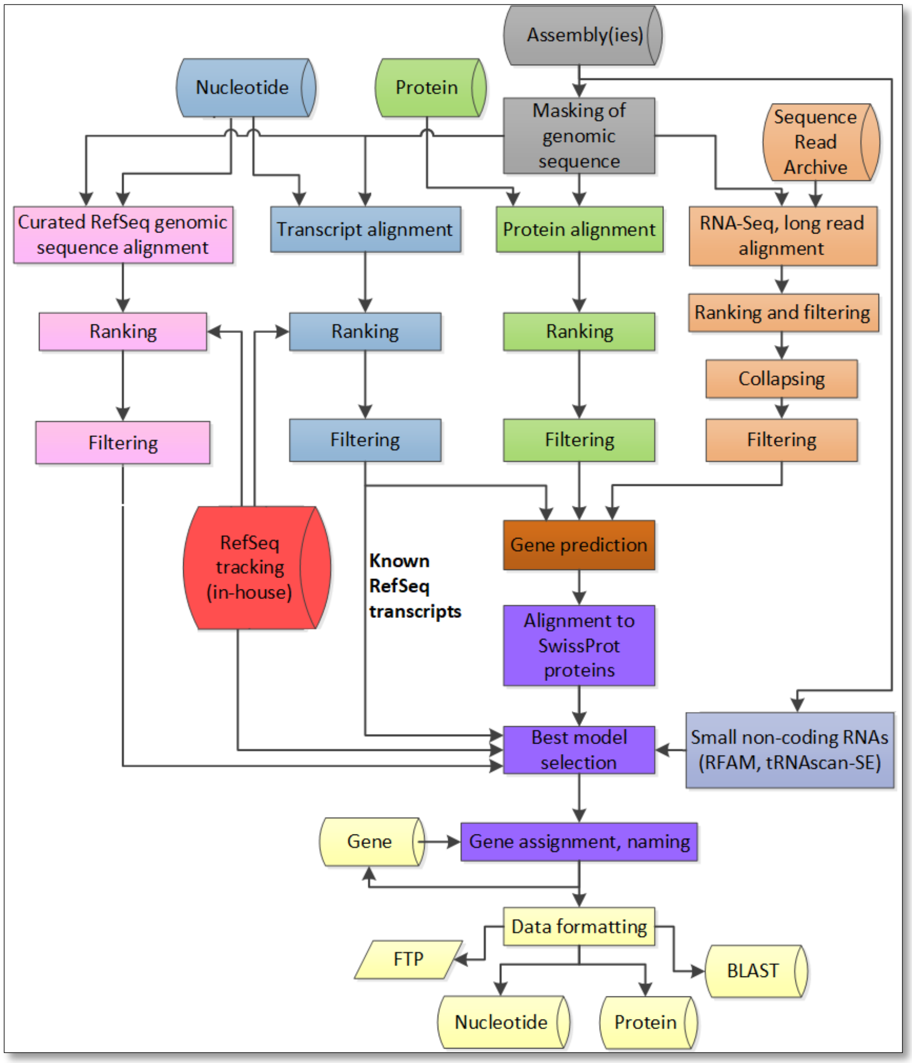
Overview

The NCBI Eukaryotic Genome Annotation Pipeline provides content for various NCBI resources including the Nucleotide, Protein, and Gene databases, the BLAST sequence alignment services, and GDV, the Genome Data Viewer. The pipeline uses a modular framework for the execution of all annotation tasks from the fetching of primary and curated data from public repositories (NCBI sequence and Assembly databases) to the alignment of sequences, the prediction of genes, the generation of annotated genomic sequence records, and finally the submission of the accessioned annotation products to public databases. Core components of the pipeline are alignment programs (Splign and ProSplign) and an HMM-based prediction program (Gnomon) developed at NCBI. Important features of this annotation pipeline include:

- Flexibility and speed
- Reliance on experimental data, in particular RNA-Seq data
- Production of models that compensate for assembly issues
- Tracking of gene loci from one annotation to the next

The flow-chart (right) provides an overview of the annotation process. The genomic sequences are masked (grey) and transcripts (blue), proteins (green), RNA-Seq reads and long SRA reads (orange) are aligned to the genome. If available for the organism being annotated, curated RefSeq genomic sequences are also aligned (pink). Gene model prediction based on transcript and protein alignments is then performed (brown). Small non-coding RNAs (sncRNAs) are predicted using tRNAscan-SE for tRNAs. rRNAs, snoRNAs and snRNAs are annotated by searching eukaryotic RFAM HMMs against the genome with Infernal's *cmsearch* (lavender). The best models are selected among the RefSeq and the predicted models, named and accessioned (purple). Finally, the annotation products are formatted and deployed to public resources (yellow).

See more information at: go.usa.gov/xp7Ey



Data Access

The products of an annotation run (chromosomes, scaffolds and model transcripts and proteins) are labeled with an Annotation Release number. The Annotation Release name is the combination of the organism name and Annotation Release number (e.g. NCBI Papio anubis Annotation Release 104) and is used throughout NCBI as a way to uniquely identify annotation products originating from the same annotation run.

The data produced for each Annotation Release is available

- In Entrez Nucleotide and Protein
- For download from the assembly FTP site
- In the Gene database
- In GDV for browsing
- In Blast databases, including nr

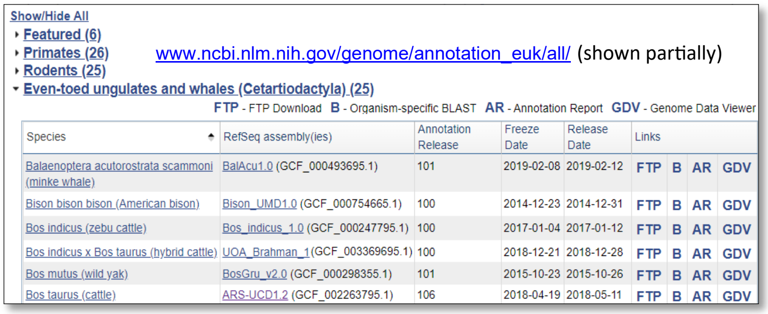
A web report summarizing the annotation results and listing the inputs used for the annotation is available for each Annotation Release. See for example: www.ncbi.nlm.nih.gov/genome/annotation_euk/Papio_anubis/104/

Annotated Organisms

Only genomes with assemblies that are public in INSDC (DDBJ, ENA or GenBank) are considered for annotation by the Eukaryotic Genome Annotation Pipeline. NCBI makes this selection based on several factors. These include:

- NIH/NCBI priorities
- Assembly quality
- Biological, evolutionary, or economic importance
- Public availability of supporting transcript evidence
- Community interest

See the full list of annotated organisms online (right).
Request an annotation at: go.usa.gov/xpsEJ



Show/Hide All
Featured (6)
Primates (26)
Rodents (25)
Even-toed ungulates and whales (Cetartiodactyla) (25)
FTP - FTP Download B - Organism-specific BLAST AR - Annotation Report GDV - Genome Data Viewer

www.ncbi.nlm.nih.gov/genome/annotation_euk/all/ (shown partially)

Species	RefSeq assembly(ies)	Annotation Release	Freeze Date	Release Date	Links			
Balaenoptera acutorostrata scammoni (minke whale)	BalAcu1.0 (GCF_000493695.1)	101	2019-02-08	2019-02-12	FTP	B	AR	GDV
Bison bison bison (American bison)	Bison_UMD1.0 (GCF_000754665.1)	100	2014-12-23	2014-12-31	FTP	B	AR	GDV
Bos indicus (zebu cattle)	Bos_indicus_1.0 (GCF_000247795.1)	100	2017-01-04	2017-01-12	FTP	B	AR	GDV
Bos indicus x Bos taurus (hybrid cattle)	UOA_Brahman_1(GCF_003369695.1)	100	2018-12-21	2018-12-28	FTP	B	AR	GDV
Bos mutus (wild yak)	BosGru_v2.0 (GCF_000298355.1)	101	2015-10-23	2015-10-26	FTP	B	AR	GDV
Bos taurus (cattle)	ARS-UCD1.2 (GCF_002263795.1)	106	2018-04-19	2018-05-11	FTP	B	AR	GDV

Input Data Sets

Source of genome assemblies: The RefSeq assemblies annotated by NCBI are copies of the genome assemblies publicly available from DDBJ, ENA and GenBank. Details of a specific assembly can be found in the Assembly database (www.ncbi.nlm.nih.gov/assembly/).

Transcripts: The set of transcripts selected for alignment to the genome varies by species. It is constrained by what is available in the public databases (Nucleotide, Protein, dbEST) for the species and closely-related organisms. It may include:

- Curated RefSeq transcripts, if available
- Full-length cDNAs and ESTs

Next generation transcriptomic data: runs of high-throughput sequences available in SRA for the annotated species or closely related species.

- RNA-Seq data
- Long read data (PacBio IsoSeq, Oxford Nanopore Technologies transcriptomes)

Proteins: Like transcripts, the set of proteins selected for alignment to the genome varies by species. It most often contains proteins from other organisms. It may include:

- Curated RefSeq proteins
- Model RefSeq proteins annotated on high quality genomes and supported by experimental evidence
- GenBank proteins derived from cDNAs (not conceptual translations)

Curated RefSeq genomic sequences: For certain organisms, a special set of genomic sequences is curated. These sequences are designated with the NG_ prefix and represent either non-transcribed pseudogenes, RefSeqGene records (human only), or complex gene clusters that are difficult to annotate by automated methods. They are aligned to the genome and their features are projected on the genome.

Annotated Gene and Transcript Features

The final set of annotated features comprises the following:

- **models based on curated RefSeq entries (NM_, NR_, and NP_ prefix):** annotated based on alignments of curated same-species RefSeq
- **miRNAs (NR_ prefix)**: annotated based on alignments of transcripts imported from miRBase into RefSeq
- **rRNAs, snoRNAs and snRNAs (NR_ prefix):** annotated based on eukaryotic RFAM HMMs hits
- **models based on Gnomon predictions (XM_ and XP_ prefix)**: based on the alignments and chaining of the RNA-Seq, transcripts and protein provided on input, and in a minority of cases *ab initio* extension
- **tRNAs:** predicted using tRNAscan-SE

Gene naming and locus type selection follow a set of rules:

- Genes represented by curated RefSeq sequences inherit the information from the RefSeq sequence
- Vertebrate and plant genes represented by predicted models are named based on orthology to human and *Arabidopsis thaliana*, respectively.
  - ◊ In the absence of an ortholog, genes are named by homology to SwissProt proteins
- Most predicted models with insertions, deletions or frameshifts are labeled as pseudogenes
- Predicted models with insertions, deletions or frameshifts may still be considered coding if they have a strong unique hit to SwissProt entries or appear to be orthologs of known protein-coding genes. Their titles will be prefixed with “PREDICTED: LOW QUALITY PROTEIN”

Link to Relevant Resources

Entrez Assembly	go.usa.gov/xp7pg	All annotated genomes:	go.usa.gov/xp7pc
Downloads:	go.usa.gov/xp7px	Genome Data Viewer:	go.usa.gov/xp7p2
Reference Sequence:	go.usa.gov/xp7pb	Splign and prosplign:	go.usa.gov/xp7pa
Gnomon:	go.usa.gov/xp7pC		