Overview
The NCBI eukaryotic genome annotation pipeline provides content for various NCBI resources including Nucleotide, Protein, and Gene databases, the BLAST sequence alignment services, and the Map Viewer genome browser. The pipeline uses a modular framework for the execution of all annotation tasks from fetching of primary and curated data from public repositories (NCBI sequence and Assembly databases) to the alignment of sequences, gene prediction, and generation of annotated genomic sequence records, to 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 by staff at NCBI. Important features of this annotation pipeline include:

- higher weight given to curated evidence than non-curated evidence
- production of models that compensate for assembly issues
- tracking of gene loci from one annotation to the next
- ability to co-annotate multiple assemblies for the same organism
- flexibility and speed

The flow-chart (right) illustrates the workflow of this pipeline. Here information and/or data retrieved from NCBI databases are represented by cylinders (A). Other input datasets, such as Transcripts and Proteins (B), are presented by rectangles. Specific processes in this interconnected workflow, such as Model prediction (C), are marked by round cornered rectangles. The final products generated by this workflow are distributed to other resources and services (D).

Data Access
The data produced by this pipeline is available from the genome ftp site and other resources such as Nucleotide, Protein and Gene databases, MapViewer, as well as BLAST. For more information and direct links to these resources, see: https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/.

Recent Changes
Annotation Release numbers are now used to differentiate the independent notions of genome assemblies and genome annotations generated here at NCBI. Annotation Release numbers:

- are integers that increment each time the genome annotation is updated.
- have initial values starting at 100 or higher
- are incremented independently for each organism, thus different organisms may have the same release number even though they were annotated at different times
- are used for the set of annotations calculated on one or more genome assemblies, if multiple assemblies were used as input

Annotation release numbers are reported in Annotation reports, on annotated sequence records in the COMMENT section, and in Map Viewer. An example is the RefSeq chromosome 1 from the human GRCh38 reference assembly (https://www.ncbi.nlm.nih.gov/nuccore/NC_000001.11) and Annotation release 108 (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/Homo_sapiens/108/).
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 (https://www.ncbi.nlm.nih.gov/assembly/). Repeat sequences present in the input assemblies are masked using Repeat-Masker or WindowMasker (for organisms without well-characterized repeats).

Transcripts: The set of transcripts selected for alignment to the genome varies by species, is limited by what is available from the public databases, and may contain transcripts from other related organisms. This set includes:
- Curated RefSeq transcripts
- Other transcripts
  - GenBank transcripts from the taxonomically-relevant GenBank divisions or batch divisions such as Third-Party Annotation, High-throughput cDNA, Transcriptome Shotgun Assembly and Expressed sequence tag
  - Long RNAseq sequences from Sequence Reads Archive

Proteins: Like transcripts, the set of proteins selected for alignment to the genome also varies by species. It may contain proteins from other organisms. It generally includes:
- Curated RefSeq proteins
- GenBank proteins derived from cDNAs from the taxonomically relevant GenBank divisions

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 a complex gene cluster that is difficult to annotate by automated methods. They are aligned to the genome during the process to identify their best placement.

Choosing the Best Models, Naming, and Assigning GeneIDs

The final set of annotated features comprises the following (in order of preference):
- models based on curated RefSeq entries: Gene annotation based on alignments of curated same-species RefSeq transcripts to genomic assembly.
- models based on Gnomon predictions: For loci without RefSeq transcript alignment, Gnomon predictions are evaluated based on a set of criteria
- miRNAs: Entries are imported from miRBase, assigned accessions with NR_prefix, and placed by Splign
- tRNAs: Loci are predicted using tRNAscan-SE

Gene naming and locus type selection follow a set of rules. Representative cases are as follows:
- Genes represented by curated RefSeq sequences inherit the information from the RefSeq sequence
- Genes represented by predicted models are named based on 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 as 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”
- When multiple genome assemblies are annotated for the same organism, a partial or imperfect model on one assembly may still be called coding if a complete model exists at the corresponding locus on another annotated assembly.

Genes in the final set of models are assigned GeneIDs in NCBI's Gene database. GeneIDs are carried forward from one annotation run to the next so they are stable in most cases. Specifically:
- A gene represented by a curated RefSeq transcript will receive the GeneID of the RefSeq transcript
- All alternative splice forms of a gene get the same GeneID
- Genes in corresponding loci of co-annotated assemblies are assigned the same GeneIDs

Link to Relevant Resources

tRNAScan-SE (outside NCBI): http://selab.janelia.org/tRNAscan-SE/