The NCBI Eukaryotic Genome Annotation Pipeline

Last

The NCBI Eukaryotic Genome Annotation Pipeline provides content for various NCBI resources including Nucleotide, Protein, BLAST, Gene and the Genome Data Viewer genome browser.

This page provides an overview of the annotation process. Please refer to the Eukaryotic Genome Annotation chapter of the NCBI Handbook for algorithmic details.

The pipeline uses a modular framework for the execution of all annotation tasks from the fetching of raw and curated data from public repositories (sequence and Assembly databases) to the alignment of sequences and the prediction of genes, 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 gene prediction program (Gnomon) developed at NCBI.

Important features of the pipeline include:

- flexibility and speed
- higher weight given to curated evidence than non-curated evidence
- utilization of RNA-Seq for gene prediction
- 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

The products of an annotation run (chromosome, 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 *Pongo abelii* Annotation Release 103) and is used throughout NCBI as a way to uniquely identify annotation products originating from the same annotation run.

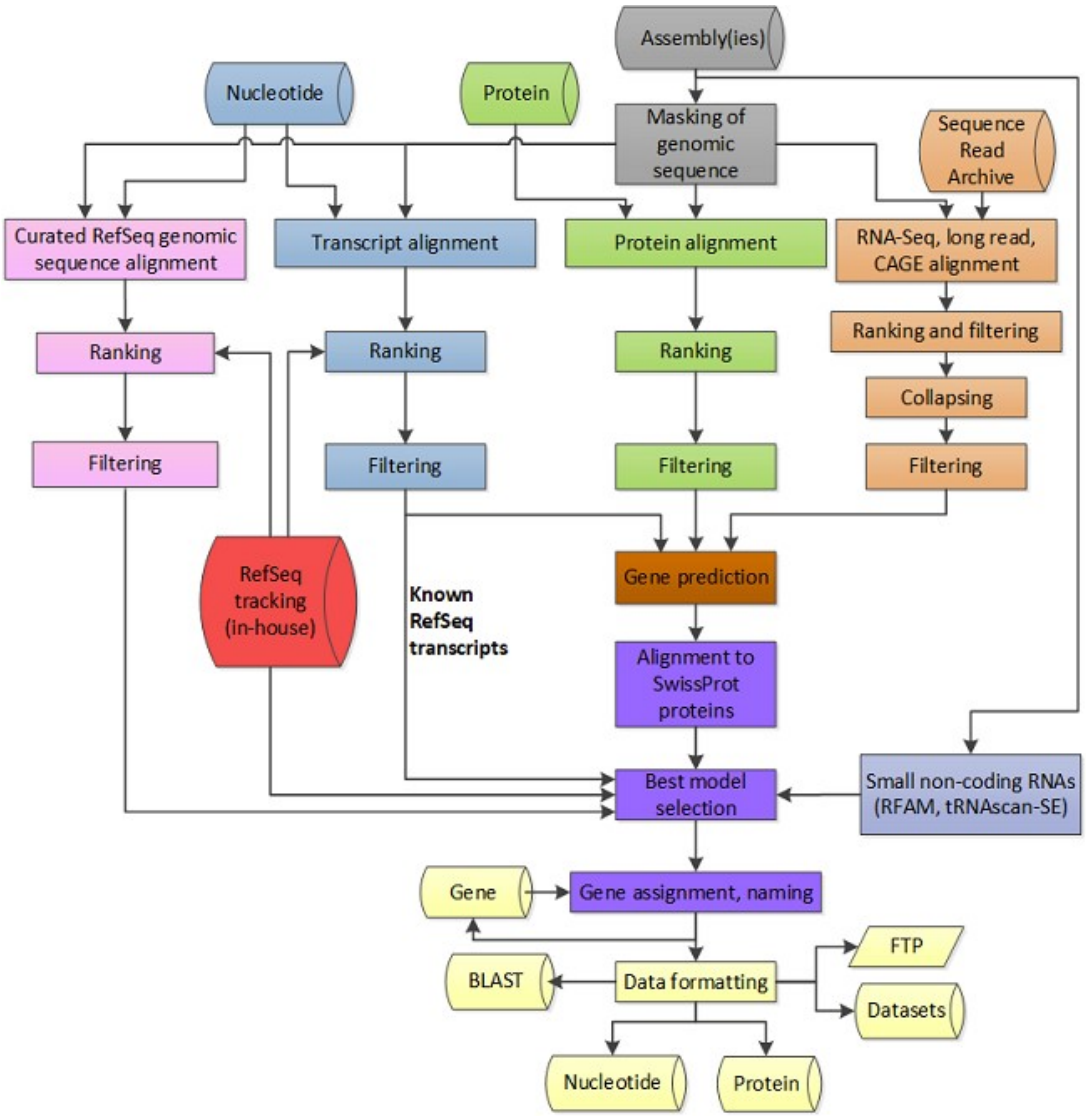
Contents

- **Process**
  - Source of genome assemblies
  - Masking
  - Transcript alignments
  - Transcriptomics long read alignments
  - RNA-Seq read alignments
  - Protein alignments
  - Model prediction
  - Curated RefSeq genomic sequence alignments
  - Choosing the best models for a gene
  - Protein naming and determination of locus type
  - Assignment of GeneIDs
  - Annotation of small RNAs
  - Annotation of transcription start sites (TSS)
- **Special considerations**
  - Annotation of multiple assemblies
  - Re-annotation
  - Annotation quality
- **Annotation products**
- **Data availability**
- **References**

Please see The Eukaryotic Genome Annotation chapter in the NCBI Handbook for more details about the algorithms.

Process

The figure below provides an overview of the annotation process. The genomic sequences are masked (grey) and transcripts (blue), proteins (green) and RNA-Seq reads and, if available in SRA, long reads transcriptomes and Cap Analysis Gene Expression (CAGE) data (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). 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).



Source of genome assemblies

The RefSeq assemblies that are annotated by NCBI are copies of the genome assemblies that are public in INSDC (DDBJ, ENA and GenBank). Unplaced scaffolds with length below 1000 bases may not be included in the RefSeq copy of the assembly if the INSDC assembly contains more than 300,000 unplaced scaffolds and more than 25,000 of them are below 1000 bases. Both RefSeq and GenBank assemblies are further described in the Assembly resource.

Masking

Masking is done using RepeatMasker or WindowMasker. Human and mouse are masked with RepeatMasker using their respective Dfam libraries, while genomes from other species are masked with WindowMasker.

Transcript alignments

The set of transcripts selected for alignment to the genome varies by species, and may include transcripts from other organisms. This set generally includes:

- Known RefSeq transcripts: Coding and non-coding RefSeq transcripts with NM_ or NR_ prefixes, respectively, are generated by NCBI staff based on automatic processes, manual curation, or data from collaborating groups (see more details here)
- GenBank transcripts from the taxonomically relevant GenBank divisions, and the Third-Party Annotation (TPA), High-throughput cDNA (HTC) and Transcriptome Shotgun Assembly (TSA) divisions
- ESTs from dbEST

Sequences highly likely to be mitochondrial or to have cloning vector or IS element contamination, and sequences identified as low quality by RefSeq curation staff are screened out.

RefSeq transcripts and non-RefSeq transcripts that pass the contamination screen are aligned locally to the genome using BLAST to identify the location(s) at which transcripts align. Global re-alignment at these locations is performed with Splign to refine the identification of splice sites. Alignments are then ranked and filtered based on customizable criteria (such as coverage, identity, rank). Typically, only the best-placed (rank 1) alignment for a given query is selected for use in the downstream steps.

Transcriptomics long read alignments

Transcriptomics reads from SRA generated using long read sequencing technologies such as PacBio or Oxford Nanopore are aligned to the genome using Minimap2. Each transcript's best-placed (rank 1) alignment is selected for use in the downstream steps, if above 85% identity.

RNA-Seq read alignments

RNA-Seq reads for the species or closely related species are aligned to the genome. When a very large number of samples amd reads (multiple billions) are available in SRA, projects with samples spanning the widest range of tissues and developmental stages are chosen over others, with a preference for untreated or non-diseased samples. RNA-Seq reads are aligned to the genome with STAR. To address the short length, redundancy and abundance of the reads, alignments with the same splice structure and the same or similar start and end points are collapsed into a single representative alignment. Information is recorded about the samples and number of reads represented by each alignment, so the level of support can be used to filter alignments and evaluate gene predictions. Alignments representing very rare introns likely to be background noise are filtered out.

Protein alignments

The set of proteins selected for alignment to the genome varies by species, and may include proteins from other organisms. This set generally includes:

- Known RefSeq proteins
- GenBank proteins derived from cDNAs from the taxonomically relevant GenBank divisions

Highly repetitive sequences are removed from the set. Proteins are aligned locally to the genome with BLAST and re-aligned globally using ProSplign. Alignments are then ranked and filtered based on customizable criteria.

Model prediction

Protein, transcript, transcriptomics and RNA-Seq read alignments are passed to Gnomon for gene prediction. Gnomon first chains together non-conflicting alignments into putative models. In a second step, Gnomon extends predictions missing a start or a stop codon or internal exon(s) using an HMM-based algorithm. Gnomon additionally creates pure *ab initio* predictions where open reading frames of sufficient length but with no supporting alignment are detected.

This first set of predictions is further refined by alignment against a subset of the nr (non-redundant) database of protein sequences. The additional alignments are added to the initial alignments, and the chaining and *ab initio* extension steps are repeated. The results constitute the set of Gnomon predictions.

Gnomon predictions may include deletions or insertions of Ns with respect to the genomic sequence. These differentes are introduced to compensate for frameshifts or stop codons in the literal translation of the genome, when the aligning proteins provides evidence of an intact ORF.

Curated RefSeq genomic sequence alignments

For some organisms, a set of genomic sequences is curated (RefSeq accessions with NG_ prefixes). These sequences represent either non-transcribed pseudogenes, a manually annotated gene cluster that is difficult to annotate via automated methods, and human RefSeqGene records. They are aligned to the genome, and their best placement is identified.

Choosing the best models for a gene

The final set of annotated features comprises, in order of preference, pre-existing RefSeq sequences and a subset of well-supported Gnomon-predicted models. It is built by evaluating together at each locus the known RefSeq transcripts, the features projected from curated RefSeq genomic alignments and the models predicted by Gnomon.

1. Models based on known and curated RefSeq

RefSeq transcripts are given precedence over overlapping Gnomon models with the same splice pattern. Alignments of known same-species RefSeq transcripts or curated genomic sequences are used directly to annotate the gene, RNA and CDS features on the genome. Since the RefSeq sequence may not align perfectly or completely to the genomic sequence, a consequence of this rule is that the annotated product may differ from the conceptual translation of the genome. Differences between the RefSeq transcripts and the genome are provided in a note on the RefSeq genomic record (scaffold or chromosome).

2. Models based on Gnomon predictions

Gnomon predictions are included in the final set of annotations if they do not share all splice sites with a RefSeq transcript and if they meet certain quality thresholds including:

- Only fully- or partially-supported Gnomon predictions, or pure *ab initio* Gnomon predictions with high coverage hits to UniProtKB/SwissProt proteins are selected
- When multiple fully-supported transcript variants are predicted for a gene, only the Gnomon predictions supported in their entirety by a single long alignment (e.g. a full-length mRNA) or by RNA-Seq reads from a single BioSample are selected
- Poorly-supported Gnomon predictions conflicting with better-supported models annotated on the opposite strand are excluded from the final set of models
- Gnomon predictions with high homology to transposable or retro-transposable elements are excluded from the final set of models

3. Integrating RefSeq and Gnomon annotations

As a result of the model selection process, a gene may be represented by multiple splice variants, with some of them known RefSeq and others model RefSeq (originating from Gnomon predictions).

Gnomon predictions selected for the final annotation set are assigned model RefSeq accessions with XM_ or XR_ prefixes for transcripts and XP_ prefixes for proteins to distinguish them from known RefSeq with NM_/NR_ and NP_ prefixes. Model RefSeq can be searched in Entrez with the query "srcdb_refseq_model[properties]" while known RefSeq sequences can be obtained with the query "srcdb_refseq_known[properties]".

Protein naming and determination of locus type

- Genes represented by known or curated RefSeq sequences inherit the Gene symbol, name and locus type (e.g. coding, pseudogene...) of the RefSeq sequence.
- Genes represented by predicted models are named based on homology to SwissProt proteins.
- Most Gnomon models with insertions, deletions or frameshifts are labeled pseudogenes.
- Gnomon models with insertions or deletions relative to the genome may be considered coding if they have a strong unique hit to the SwissProt database or appear to be orthologs of known protein-coding genes. Titles for these models are prefixed with "PREDICTED: LOW QUALITY PROTEIN" to indicate that these models and the underlying assembly sequences may content defects.
- Gnomon models that appear to be single-exon retrocopies of protein-coding genes may be annotated as pseudogenes.
- When multiple assemblies are annotated, a partial or imperfect model may be called coding because a complete model exists at the corresponding locus on one of the other annotated assemblies.

Assignment of GeneIDs

Genes in the final set of models are assigned GeneIDs in NCBI's Gene database.

- A gene represented by a known RefSeq transcript will receive the GeneID of the RefSeq transcript.
- All alternative splice forms of a gene get the same GeneID.
- As much as possible, GeneIDs are carried forward from one annotation run to the next, using the mapping of the new assembly to the previous one if the assembly was updated.
- Gene features mapped to equivalent locations of co-annotated assemblies are assigned the same GeneIDs.

Annotation of small RNAs

- miRNAs are imported from miRBase, accessioned with NR_ prefixes and placed using Splign.
- tRNAs are predicted with tRNAscan-SE.
- Starting with software version 8.0, rRNAs, snoRNAs and snRNAs are annotated by searching eukaryotic RFAM HMMs against the genome with Infernal's *cmsearch*.

Annotation of transcription start sites (TSS)

Starting with software release 9.0, Cap Analysis Gene Expression (CAGE) data that is available in SRA for the species are aligned to the genome with Splign and used for annotating transcription start sites.

Special considerations

Annotation of multiple assemblies

When multiple assemblies of good quality are available for a given organism, annotation of all is done in coordination. To ensure that matching regions across assemblies are annotated the same way, assemblies are aligned to each other before the annotation.

- Assembly-assembly alignment results are used to rank the transcript and the curated genomic alignments: for a given query sequence, alignments to corresponding regions of two assemblies receive the same rank.
- Corresponding loci of multiple assemblies are assigned the same GeneID and locus type.

Assembly-assembly alignments are available through the NCBI Genome Remapping Service.

Re-annotation

Organisms are periodically re-annotated when new evidence is available (e.g. RNA-Seq) or when a new assembly is released. Special attention is given to tracking of models and genes from one release of the annotation to the next. Previous and current models annotated at overlapping genomic locations are identified and the locus type and GeneID of the previous models are taken into consideration when assigning GeneIDs to the new models. If the assembly was updated between the two rounds of annotation, the assemblies are aligned to each other and the alignments used to match previous and current models in mapped regions.

Annotation quality

The quality of the annotation is assessed prior to publishing, based on the intrinsic characteristics of the annotated models and on the expectations for the species. Indicators of a low quality annotation may disqualify a genome from being included in RefSeq. These indicators are: high count of coding genes that lack near-full coverage by alignments of experimental evidence, high count of partial coding genes (lacking a start or stop codon, or internal exons), high count of low-quality genes with suspected frameshifts or premature stop codons, low BUSCO completeness score (see below), and, for vertebrates, low count of genes with orthologs to a reference species.

BUSCO run in "protein" mode provides an estimate of the completeness of the gene set. The BUSCO models (single-copy marker genes) for the most fitting lineage based on NCBI Taxonomy are searched against the longest protein for each annotated coding gene. Results are reported in BUSCO notation (C:complete [S:single-copy, D:duplicated], F:fragmented, M:missing, n:number of genes used).

Annotation products

- The products of the annotation process comprise:
  - The scaffolds and chromosomes of the assembled genomes, with the annotation products as features.
  - The individual products (transcripts and proteins)

Product	Origin of the product	Note for the features on the scaffolds and chromosomes*
Known transcripts/proteins (NM_, NR_, NP_)	curated RefSeq genomic	"Derived by automated computational analysis using gene prediction method: Curated Genomic"
Known transcripts/proteins (NM_, NR_, NP_)	known RefSeq transcript	"Derived by automated computational analysis using gene prediction method: BestRefseq"
Model transcripts/proteins (fully or partially supported) (XM_, XR_, XP_)	Gnomon	"Derived by automated computational analysis using gene prediction method: Gnomon"
Model short non-coding transcripts (XR_)	Rfam + cmsearch	"Derived by automated computational analysis using gene prediction method: cmsearch"
tRNAs (no accession)	tRNAscan-SE	"tRNA features were annotated by tRNAscan-SE"
Non-transcribed pseudogenes (no accession)	curated RefSeq genomic	"Derived by automated computational analysis using gene prediction method: Curated Genomic"
Non-transcribed pseudogenes (no accession)	Gnomon	"Derived by automated computational analysis using gene prediction method: Gnomon"

Product	Origin of the product	Note for the features on the scaf and chromosomes*
Full set of Gnomon predictions (no accession)	Gnomon	NA. Not in the sequence database. Av on the FTP site and as BLAST databas

* For predicted models, the note is also on the records of individual annotation products.

- Sequence records for predicted models, scaffolds and chromosomes contain the Annotation Release number, which in combination with the species uniquely identifies the annotation. For example, the sequence records for scaffolds, chromosomes and predicted transcripts and proteins for NCBI *Pongo abelii* Annotation Release 103 contain the following comment:

```
##Genome-Annotation-Data-START##
Annotation Provider       :: NCBI
Annotation Status         :: Full annotation
Annotation Name           :: Pongo abelii Annotation Release 103
Annotation Version        :: 103
Annotation Pipeline       :: NCBI eukaryotic genome annotation pipeline
Annotation Software Version :: 8.0
Annotation Method         :: Best-placed RefSeq; Gnomon
Features Annotated        :: Gene; mRNA; CDS; ncRNA
##Genome-Annotation-Data-END##
```

Data availability

The data produced by the annotation pipeline is available in various resources:

- Nucleotide
- Protein
- BLAST
- Gene
- Genome Data Viewer
- FTP site

References

- BUSCO: Manni M et al. Molecular biology and evolution 2021, **38**(10):4647-4654
- Minimap2: Li H. *Bioinformatics* 2018 **34**(18):3094-3100
- miRBase: Griffiths-Jones S. *Nucleic Acids Research* 2004, **32**(Database Issue):D109-11
- RefSeq: Pruitt KD et al. *Nucleic Acids Research* 2014, **42**(Database issue):D756-63
- RepeatMasker: Smit AFA, Hubley R, Green P. RepeatMasker Open-3.0. 1996–2004. http://www.repeatmasker.org
- Rfam: Nawrocki, EP et al. *Nucleic Acids Research* 2015, **43**(Database issue):D130-7
- Splign: Kapustin Y, Souvorov A, Tatusova T, Lipman D. *Biology Direct* 2008, **3**:20
- STAR: Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. *Bioinformatics* 2013, **29**(1): 15–21
- tRNAscan-SE: Lowe, TM and Eddy, SR. *Nucleic Acids Research* 1997, **25**: 955-964
- WindowMasker: Morgulis A, Gertz EM, Schäffer AA, Agarwala R.*Bioinformatics* 2006 **2**:134-41