Annotation of Laurasiatheria genomes by National Center for Biotechnology Information (NCBI)

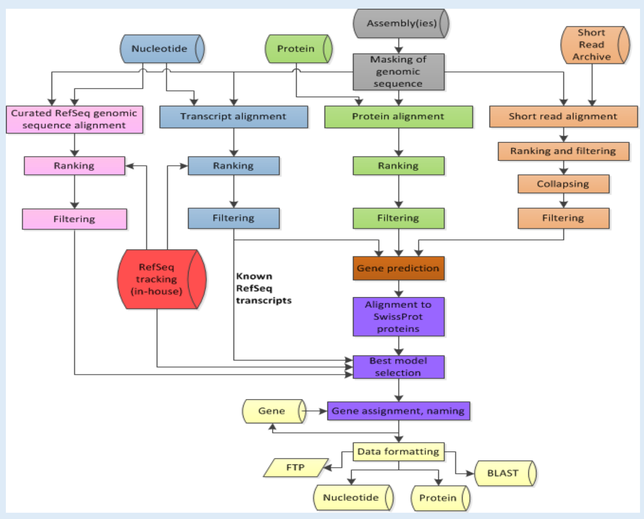
Shashikant Pujar, Bhanu Rajput, Francoise Thibaud-Nissen, Terence D. Murphy, Paul Kitts, Kim D. Pruitt
National Center for Biotechnology Information (NCBI), National Institutes of Health, Bethesda MD

Abstract

National Center for Biotechnology Information's (NCBI) Eukaryotic Genome Annotation (http://www.ncbi.nlm.nih.gov/genome/annotation_euk/) generates comprehensive sets of genes, transcripts and proteins for finished and unfinished public genome assemblies of a wide range of taxa among invertebrates, vertebrates and plants. Among them is Laurasiatheria, a sub-group of placental mammals that includes several farm and barnyard animals, carnivores as well as moles, shrews, bats, whales and dolphins. NCBI has annotated 35 genomes from this mammalian group till date, many of which incorporate improved features of the annotation pipeline. One such feature is the intelligent transcript model generation utilizing transcript data from GenBank, to circumvent genome assembly gaps. A second feature is the use of RNA_seq data for transcript model prediction (in horse, for example).

These genomes are in scope for curation support from NCBI's RefSeq group, which results in addition of "known" RefSeqs (NG_, NM_/NP_, NR_ accessions). These RefSeqs, in turn, serve as reagents for the annotation of related species/taxa that have little or no transcript data. In addition, the genomes also benefit from curation of genes that exhibit exceptional biology. RefSeqs for such genes are annotated across taxa with the transcript region involved in exceptional biology flagged as a miscellaneous feature on the RefSeq record. For example OAZ1 (translational frameshift) is represented in pig (NM_001122994.2), dog (NM_001127234.1) and cow (NM_001127243.2), among others. The streamlined NCBI annotation pipeline, combined with RefSeq curation support, are equipped to provide reliable and value-added annotations to genomes of your interest.

NCBI'S EUKARYOTIC ANNOTATION PIPELINE

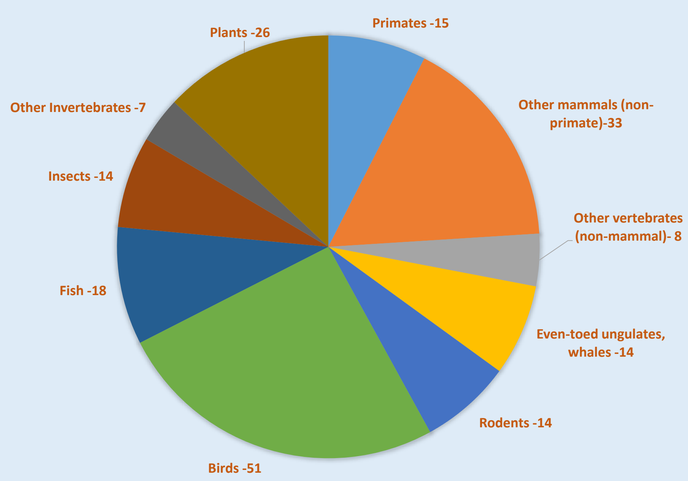


- **Reagents**: RefSeqs, GenBank transcripts and proteins; RNA-seq data
- **Alignments**: Splign and ProSplign
- **Best Alignments**: Filter to obtain best alignments
- **Predict Models**: Gnomon predicts gene models, best models added to RefSeq set (XM_/XP, XR_)
- **Ortholog calls**: Based on reciprocal best hit and synteny calculation
- **Assign gene IDs and names**: Orthology used to assign names and locus types

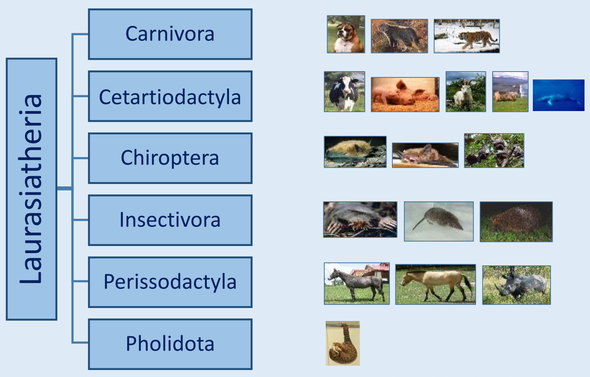
Key features of annotation

- Assemblies in INSDC are in scope.
- Annotations cover a broad range of taxonomic groups.
- Over 100 assemblies annotated in 2014.
- Speed: 5-10 days.
- Use of RNA-seq data in annotation.
- Corrected model prediction using base correction and gap-filling using transcripts.
- Enhanced public reporting (annotation reports).

CURRENTLY ANNOTATED EUKARYOTIC GENOMES



Laurasiatheria sub-groups and representative genomes



Laurasiatheria genomes at NCBI

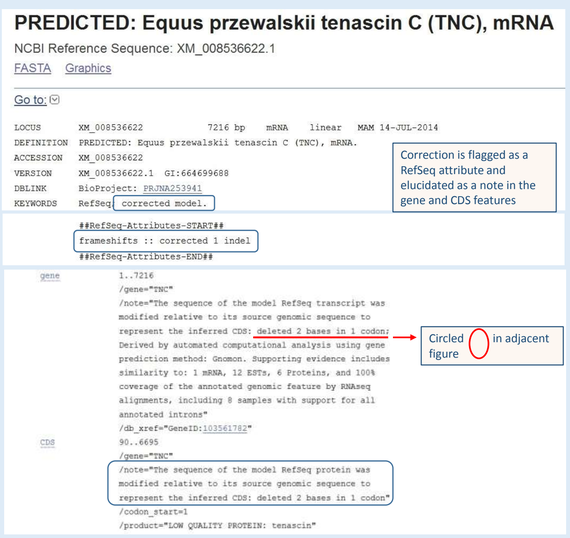
- 35 genomes annotated and represented in Gene
- 19 genomes annotated using RNA-seq data
- **Most recent annotation**
  Camelus ferus (wild Bactrian camel)
- **Coming soon!**
  Bison bison bison (American bison)
  Camelus bactrianus (Bactrian camel)
  Camelus dromedarius (Arabian camel)
  Bos taurus (cow: re-annotation)

Annotation details of selected Laurasiatheria genomes

Name	Common name	Assembly last annotated	Current Annotation	Total genes*	Total RefSeq RNA**
Carnivora					
Canis lupus familiaris	dog	CanFam3.1	Annotation Release 103	30,524	48,936
Felis catus	cat	Felis_catus-6.2	Annotation Release 101	24,798	32,587
Panthera tigris altaica	Amur tiger	PanTig1.0	Annotation Release 100	24,653	30,314
Cetartiodactyla					
Bos taurus	cow	Bos_taurus_UMD_3.1	Annotation Release 103	30,777	55,833
Bos taurus	cow	Btau_4.6.1	Annotation Release 103	32,517	36,964
Sus scrofa	pig	Sscrofa10.2	Annotation Release 104	47,292	51,530
Capra hircus	goat	CHIR_1.0	Annotation Release 100	24,657	29,755
Ovis aries	sheep	Oar_v3.1	Annotation Release 100	24,102	23,095
Balaenoptera acutorostrata scammoni	Scammon's minke whale	BalAcu1.0	Annotation Release 100	24,099	37,872
Chiroptera					
Myotis lucifugus	little brown bat	Myoluc2.0	Annotation Release 100	26,610	30,523
Eptesicus fuscus	big brown bat	EptFus1.0	Annotation Release 100	23,121	24,620
Pteropus alecto	black flying fox	ASM32557v1	Annotation Release 100	21,449	23,635
Insectivora					
Condylura cristata	star-nosed mole	ConCri1.0	Annotation Release 100	19,708	22,724
Sorex araneus	European shrew	SorAra2.0	Annotation Release 100	22,717	22,846
Erinaceus europaeus	western European hedgehog	EriEur2.0	Annotation Release 100	22,146	24,357
Perissodactyla					
Equus caballus	domestic horse	EquCab2.0	Annotation Release 101	25,768	36,435
Equus przewalskii	wild horse	Burgud	Annotation Release 100	28,179	45,076
Ceratotherium simum simum	southern white rhinoceros	CerSimSim1.0	Annotation Release 100	22,175	26,490

*per Entrez Gene query: txid####[orgn] AND alive[property] **per Entrez Nucleotide query: "Scientific name"[Organism] AND biomol_rna[prop] AND srcdb_refseq[prop]

CORRECTED REFSEQ MODELS BASED ON INPUT TRANSCRIPT/PROTEIN DATA



Corrected and gap-filled model RefSeqs

- When the aligned transcripts differ from the genome assembly, corrections are made in predicted models, versus the genomic sequence, to maintain the integrity of the inferred protein.
- Types of corrections include insertion/deletion of bases to avoid a frameshift and base substitution to avoid an internal stop codon.
- Transcript sequence is used in model RefSeqs to circumvent assembly gaps.

Alignment of the wild horse tenascin-C model transcript XM_008536622.1 (nucleotides 3264 to 3287) to *E. przewalskii* genomic sequence (scaffold NW_007673021.1, nucleotides 593322 to 593355). Alternate codons are shaded. Amino-acids encoded by the codons are shown at the bottom.

```
3264   A G A C A C A A G A G C A A - - G C C T G C A C G T   3287
       | | | | | | | | | | | | |     | | | | | | | | | |
593322 A G A C A C A A G A G C A A A A G C C T G C A C G T   593355
1059     R     H     K     S     K       P     A     R     1066
```

Multiple sequence alignment of pig, cow and human tenascin-C proteins with the wild horse model protein.

```
E. przewalskii  XP_008534844   1059   R H K S K P A R   1066
S. scrofa       NP_999395      1059   R H K S K P A R   1066
B. taurus       NP_001071494   1059   R H K S K P A R   1066
H. sapiens      NP_002151      1059   R H K S K P A R   1066
```

CURATION SUPPORT FOR ANNOTATED GENOMES BY REFSEQ GROUP

RefSeq: NCBI Reference Sequence Database

A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.

Useful Contacts/Links

Contact: refseq-admin@ncbi.nlm.nih.gov (RefSeq related inquiries and comments)
pujars@ncbi.nlm.nih.gov (Laurasiatheria Genome Champion)
Ftp: ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/vertebrate_mammalian/ (Genome specific)
ftp://ftp.ncbi.nlm.nih.gov/refseq/ (RefSeq content across all taxa)

Scope of curation

- Names and aliases, locus type, gene boundary.
- Transcript variants and protein isoforms- variant descriptions, gene summary, publications, GeneRIF.
- Feature annotation (examples: polyA site and signal, signal peptide and mature peptide).
- Features that represent 'exceptional biology'.

Curation approach

Curation of these genomes is generally in response to:

- Internal QA tests that flag errors, including assembly issues that affect the RefSeq sequence.
- Focused curation by Genome Champions.
- RefSeq curation in response to user feedback.

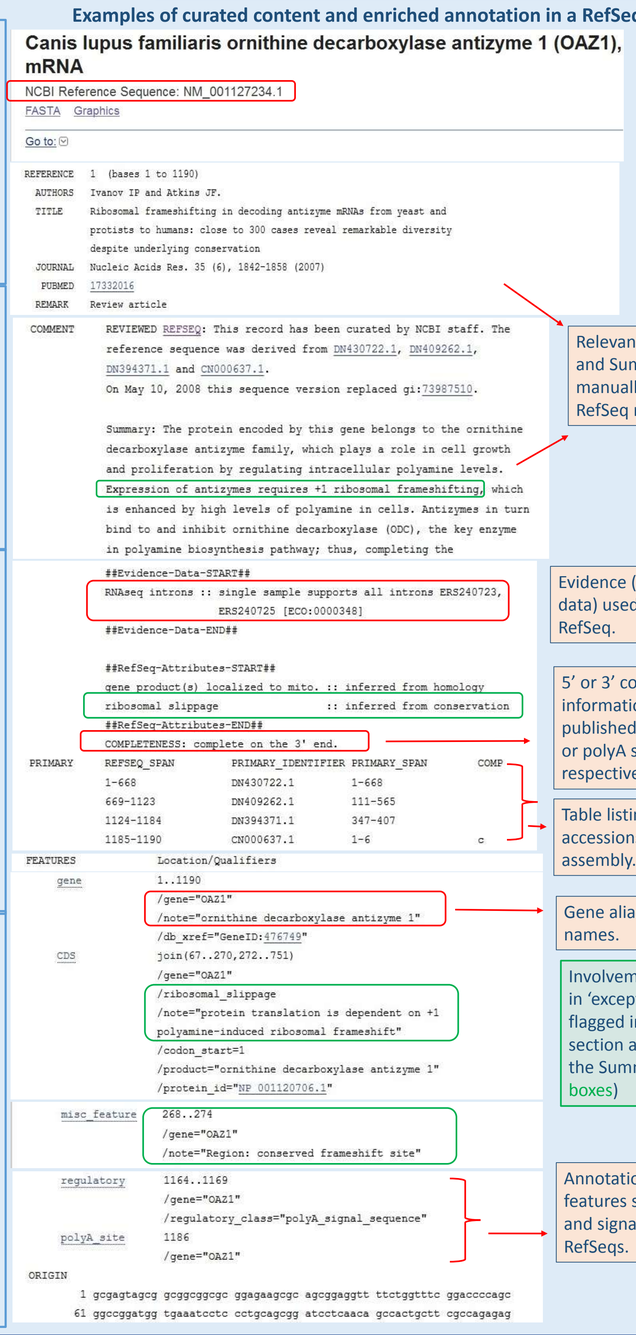
Curation by Genome Champions

- Initial setup of genome annotation and handling community requests for annotation.
- Post-annotation setup/QA of NCBI resources such as Blast page, Genome overview page and Map Viewer.
- Communication with genome communities and addressing specific genome-related queries.
- General outreach to genome communities in meetings/conferences.

Curation of 'exceptional biology' attributes

- All Laurasiatheria genomes are in scope of this specialized curation
- Computational annotation fails to annotate the features correctly, hence manual curation is needed.
- Loci curated based on literature review, annotation on primary data or by inference from conservation.
- The 'exceptional biology' feature is annotated in different sections of the RefSeq record- Summary, RefSeq attributes, CDS feature and miscellaneous feature (misc_feature).

Examples of curated content and enriched annotation in a RefSeq record



RefSeq transcripts and proteins are created based on available data

Relevant publications and Summary added manually to 'reviewed' RefSeq records.

Evidence (user-provided data) used to make the RefSeq.

5' or 3' completeness information, based on published promoter study or polyA site and signal, respectively

Table listing the GenBank accessions used in RefSeq assembly.

Gene aliases; descriptive names.

Involvement of the gene in 'exceptional biology' is flagged in 'features' section and elucidated in the Summary (green boxes)

Annotation of additional features such as polyA site and signal in 'reviewed' RefSeqs.

Acknowledgements: Routine curation support provided by curators of the RefSeq group. Several people at NCBI provide support to the eukaryotic annotation pipeline. This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.