

NGS data analysis course

RNA-Seq Quantification and Isoforms Finding

Ignacio Medina
David Montaner & Marta Bleda



PRINCIPE FELIPE
CENTRO DE INVESTIGACION

Titulo

- **National Center for Biotechnology Information (NCBI)**
<http://www.ncbi.nlm.nih.gov/>
- part of the United States National Library of Medicine (NLM)
<https://www.nlm.nih.gov/>
- a division of the National Institutes of Health (NIH)
<http://www.nih.gov/>

Resources

Databases of biomedical and genomic information for all organisms:

- Submission
 - GenBank
- Databases
 - GenBank
 - **RefSeq**
 - ...
- Downloads
 - FTP sites
 - PubMed
- Tools
 - Basic Local Alignment Search Tool (BLAST)
 - ...

See: <http://www.ncbi.nlm.nih.gov/guide/all/>

Documents

NCBI Help Manual:

- quick overview about topics
- usually just FAQs
- online (HTML) book

NCBI Handbook:

- nice introduction to each tool or database
- online (HTML) book
- but Chapters may be downloaded in PDF format
- see a [chapter example describing GenBank](#)

Many other:

- [Glossary](#)
- [NCBI Educational Resources](#)

Entrez: Search and Retrieval System

- the indexing and retrieval system used at the NCBI
- used for **all** major NCBI databases:
 - PubMed
 - Nucleotide and Protein Sequences
 - Protein Structures
 - Complete Genomes,
 - Taxonomy
 - OMIM
 - ...
- **text-based** searches over several record fields
- In practical terms, the [web interface](#)

GenBank

- collection of publicly available *annotated* nucleotide sequences and their **protein** translations
 - mRNA sequences with coding regions
 - segments of genomic DNA with single or multiple genes
 - ribosomal RNA gene clusters
 - genome shotgun reads
 - isolated genes
 - complete genomes
 - ...
- primary sequence data; **not curated**; minor checks done by the NCBI
- just authors submit and revise
- may have multiple records for same loci
- records can contradict each other
- no limit to species included

INSDC: International Nucleotide Sequence Database Collaboration

INSDC members:

- **GenBank**
- **ENA**: European Nucleotide Archive
- **DDBJ**: DNA Data Bank of Japan

<http://www.insdc.org/>

GenBank Access

<https://www.ncbi.nlm.nih.gov/genbank/>

Primarily access via the NCBI **Nucleotide** database which is divided into three divisions:

- **CoreNucleotide**: the main collection (same as Nucleotide)
- **dbEST**: single-read transcript sequences (Expressed Sequence Tags)
- **dbGSS**: unannotated short single-read primarily genomic sequences

But some other ways are available:

- **BLAST**: align against GenBank sequences
- **FPT**

GenBank record format

See an [Example of GenBank Record](#)

RefSeq: The Reference Sequence database

<http://www.ncbi.nlm.nih.gov/refseq/>

- a **curated** collection of DNA, RNA, and protein sequences
- created by the NCBI from existing data (GeneBank)
- unique example of each natural biological molecule (for each major organisms)
- not all organisms available
- for each model organism, RefSeq aims to provide separate and linked records for:
 - the genomic DNA
 - the gene transcripts
 - and the proteins arising from those transcripts

RefSeq

- non-redundant set of reference standards (**NR**)
- includes:
 - chromosomes
 - complete genomic molecules (organelle genomes, viruses, plasmids)
 - intermediate assembled genomic contigs
 - curated genomic regions, mRNAs, RNAs
 - proteins
 - alternatively spliced transcripts
- generated to provide reference standards for multiple purposes
- facilitates database inquiries based on:
 - genomic location
 - sequence
 - text annotation

RefSeq Access

- Entrez: <http://www.ncbi.nlm.nih.gov/refseq/>
- NCBI Gene: include nomenclature, maps, pathways ...
- NCBI Genome: information on genomes including sequences, maps, chromosomes, assemblies, and annotations
- NCBI Assembly: Genome assembly
- NCBI UniGene: A Unified View of the Transcriptome

Example [Homo sapiens \(human\)](#)

RefSeq record

Each RefSeq record represents a synthesis, of the primary information that was generated and submitted by many researchers.

Consistent framework between:

- sequence
- genetic
- expression
- functional information
- ...

RefSeq records are similar in format to GenBank but may include a unique accession prefix followed by

RefSeq Accession Format

Accession format: accession number that begins with two characters followed by an underscore.

There are several [RefSeq accession prefixes](#)

- NM_: mRNA
- NR_: RNA (non coding)
- NC_: Complete genomic molecule, usually a reference assembly.

Curation **VERSION** is indicated after a dot:

- NM_000014.4
- NM_000014.5

Usual fasta id for a sequence:

```
>gi|262118207|ref|NM_000202.5| Homo sapiens iduronate ...
```

Read about [Gls](#)

RefSeq Curation Levels

There are several RefSeq curation levels.

See [status codes here](#)

RefSeq records with a status of **VALIDATED** or **REVIEWED** are intended to represent the current state of genomic knowledge.

FTP Downloads

See: `ftp://ftp.ncbi.nlm.nih.gov/refseq/`

Use shell command: `wget`