Introduction to Databases
part 2

Shifra Ben-Dor
Irit Orr

BBCU

WEIZMANN INSTITUTE OF SCIENCE

Genbank Format

Divided into three parts:

- Information lines
- Feature table
- Sequence

```
LOCUS       X56734                  1859 bp    mRNA    linear   PLN 25-NOV-2005
DEFINITION  Trifolium repens mRNA for non-cyanogenic beta-glucosidase.
ACCESSION   X56734 S46826
VERSION     X56734.1  GI:21954
KEYWORDS    beta-glucosidase.
SOURCE      Trifolium repens (white clover)
  ORGANISM  Trifolium repens
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
            rosids; eurosids I; Fabales; Fabaceae; Papilionoideae; Trifolieae;
            Trifolium.
REFERENCE   1  (bases 1 to 1859)
  AUTHORS   Oxtoby,E., Dunn,M.A., Pancoro,A. and Hughes,M.A.
  TITLE     Nucleotide and derived amino acid sequence of the cyanogenic
            beta-glucosidase (linamarase) from white clover (Trifolium repens
            L.)
  JOURNAL   Plant Mol. Biol. 17 (2), 209-219 (1991)
   PUBMED   1907511
REFERENCE   2  (bases 1 to 1859)
  AUTHORS   Hughes,M.A.
  TITLE     Direct Submission
  JOURNAL   Submitted (19-NOV-1990) Hughes M.A., University of Newcastle Upon
            Tyne, Medical School, Newcastle Upon Tyne, NE2 4HH, UK
COMMENT     On Jun 10, 2005 this sequence version replaced gi:233395.
FEATURES             Location/Qualifiers
     source          1..1859
                     /organism="Trifolium repens"
                     /mol_type="mRNA"
                     /db_xref="taxon:3899"
                     /clone="TRE361"
                     /tissue_type="leaves"
                     /clone_lib="lambda gt10"
     mRNA            1..1859
                     /experiment="experimental evidence, no additional details
                     recorded"
     CDS             14..1495
                     /EC_number="3.2.1.21"
                     /note="non-cyanogenic"
                     /codon_start=1
                     /product="beta-glucosidase"
                     /protein_id="CAA40058.1"
                     /db_xref="GI:21955"
                     /db_xref="GOA:P26204"
                     /db_xref="InterPro:IPR001360"
                     /db_xref="UniProtKB/Swiss-Prot:P26204"
                     /translation="MDFIVAIFALFVISSFTITSTNAVEASTLLDIGNLSRSSFPRGF
                     IFGAGSSAYQFEGAVNEGGRGPSIWDTFTHKYPEKIRDGSNADITVDQYHRYKEDVGI
                     MKDQNMDSYRFSISWPRILPKGKLSGGINHEGIKYYNNLINELLANGIQPFVTLFHWD
                     LPQVLEDEYGGFLNSGVINDFRDYTDLCFKEFGDRVRYWSTLNEPWVFSNSGYALGTN
                     APGRCSASNVAKPGDSGTGPYIVTHNQILAHAEAVHVYKTKYQAYQKGKIGITLVSNW
                     LMPLDDNSIPDIKAAERSLDFQFGLFMEQLTTGDYSKSMRRIVKNRLPKFSKFESSLV
                     NGSFDFIGINYYSSSYISNAPSHGNAKPSYSTNPMTNISFEKHGIPLGPRAASIWIYV
                     YPYMFIQEDFEIFCYILKINITILQFSITENGMNEFNDATLPVEEALLNTYRIDYYYR
                     HLYYIRSAIRAGSNVKGFYAWSFLDCNEWFAGFTVRFGLNFVD"
```

```
ORIGIN
        1 aaacaaacca aatatggatt ttattgtagc catatttgct ctgtttgtta ttagctcatt
       61 cacaattact tccacaaatg cagttgaagc ttctactctt cttgacatag gtaacctgag
      121 tcggagcagt tttcctcgtg gcttcatctt tggtgctgga tcttcagcat accaatttga
      181 aggtgcagta aacgaaggcg gtagaggacc aagtatttgg gataccttca cccataaata
      241 tccagaaaaa ataagggatg gaagcaatgc agacatcacg gttgaccaat atcaccgcta
      301 caaggaagat gttgggatta tgaaggatca aaatatggat tcgtatagat tctcaatctc
      361 ttggccaaga atactcccaa agggaaagtt gagcggaggc ataaatcacg aaggaatcaa
      421 atattacaac aaccttatca acgaactatt ggctaacggt atacaaccat ttgtaactct
      481 ttttcattgg gatcttcccc aagtcttaga agatgagtat ggtggtttct taaactccgg
      541 tgtaataaat gattttcgag actatacgga tctttgcttc aaggaatttg gagatagagt
      601 gaggtattgg agtactctaa atgagccatg ggtgtttagc aattctggat atgcactagg
      661 aacaaatgca ccaggtcgat gttcggcctc caacgtggcc aagcctggtg attctggaac
      721 aggaccttat atagttacac acaatcaaat tcttgctcat gcagaagctg tacatgtgta
      781 taagactaaa taccaggcat atcaaaaggg aaagataggc ataacgttgg tatctaactg
      841 gttaatgcca cttgatgata atagcatacc agatataaag gctgccgaga gatcacttga
      901 cttccaattt ggattgttta tggaacaatt aacaacagga gattattcta agagcatgcg
      961 gcgtatagtt aaaaaccgat tacctaagtt ctcaaaattc gaatcaagcc tagtgaatgg
     1021 ttcatttgat tttattggta taaactatta ctcttctagt tatattagca atgccccttc
     1081 acatggcaat gccaaaccca gttactcaac aaatcctatg accaatattt catttgaaaa
     1141 acatgggata cccttaggtc caagggctgc ttcaatttgg atatatgttt atccatatat
     1201 gtttatccaa gaggacttcg agatcttttg ttacatatta aaaataaata taacaatcct
     1261 gcaattttca atcactgaaa atggtatgaa tgaattcaac gatgcaacac ttccagtaga
     1321 agaagctctt ttgaatactt acagaattga ttactattac cgtcacttat actacattcg
     1381 ttctgcaatc agggctggct caaatgtgaa gggtttttac gcatggtcat ttttggactg
     1441 taatgaatgg tttgcaggct ttactgttcg ttttggatta aactttgtag attagaaaga
     1501 tggattaaaa aggtacccta agctttctgc ccaatggtac aagaactttc tcaaaagaaa
     1561 ctagctagta ttattaaaag aactttgtag tagattacag tacatcgttt gaagttgagt
     1621 tggtgcacct aattaaataa aagaggttac tcttaacata ttttttaggcc attcgttgtg
     1681 aagttgttag gctgttattt ctattatact atgttgtagt aataagtgca ttgttgtacc
     1741 agaagctatg atcataacta taggttgatc cttcatgtat cagtttgatg ttgagaatac
     1801 tttgaattaa aagtcttttt ttatttttt aaaaaaaaaa aaaaaaaaaa aaaaaaaaa
//
```

EMBL sequence format

```
RN   [2]
RA   Wirsel S.G.R., Leibinger W., Mendgen K.W.;
RT   "Genetic diversity of fungi associated with common reed (Phragmites
RT   australis)";
RL   Unpublished.
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..581
FT                   /db_xref="taxon:112223"
FT                   /organism="ascomycota sp. 4/97-9"
FT                   /isolate="4/97-9"
```

```
ID   X56734; SV 1; linear; mRNA; STD; PLN; 1859 BP.
XX
AC   X56734; S46826;
XX
DT   12-SEP-1991 (Rel. 29, Created)
DT   25-NOV-2005 (Rel. 85, Last updated, Version 11)
XX
DE   Trifolium repens mRNA for non-cyanogenic beta-glucosidase
XX
KW   beta-glucosidase.
XX
OS   Trifolium repens (white clover)
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; rosids;
OC   eurosids I; Fabales; Fabaceae; Papilionoideae; Trifolieae; Trifolium.
XX
RN   [5]
RP   1-1859
RX   DOI; 10.1007/BF00039495.
RX   PUBMED; 1907511.
RA   Oxtoby E., Dunn M.A., Pancoro A., Hughes M.A.;
RT   "Nucleotide and derived amino acid sequence of the cyanogenic
RT   beta-glucosidase (linamarase) from white clover (Trifolium repens L.)";
RL   Plant Mol. Biol. 17(2):209-219(1991).
XX
RN   [6]
RP   1-1859
RA   Hughes M.A.;
RT   ;
RL   Submitted (19-NOV-1990) to the EMBL/GenBank/DDBJ databases.
RL   Hughes M.A., University of Newcastle Upon Tyne, Medical School, Newcastle
RL   Upon Tyne, NE2 4HH, UK
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..1859
FT                   /organism="Trifolium repens"
FT                   /mol_type="mRNA"
FT                   /clone_lib="lambda gt10"
FT                   /clone="TRE361"
FT                   /tissue_type="leaves"
FT                   /db_xref="taxon:3899"
FT   CDS             14..1495
FT                   /product="beta-glucosidase"
FT                   /EC_number="3.2.1.21"
FT                   /note="non-cyanogenic"
FT                   /db_xref="GOA:P26204"
FT                   /db_xref="HSSP:1CBG"
FT                   /db_xref="InterPro:IPR018120"
FT                   /db_xref="UniProtKB/Swiss-Prot:P26204"
FT                   /protein_id="CAA40058.1"
FT                   /translation="MDFIVAIFALFVISSFTITSTNAVEASTLLDIGNLSRSSFPRGFI
FT                   FGAGSSAYQFEGAVNEGGRGPSIWDTFTHKYPEKIRDGSNADITVDQYHRYKEDVGIMK
FT                   DQNMDSYRFSISWPRILPKGKLSGGINHEGIKYYNNLINELLANGIQPFVTLFHWDLPQ
FT                   VLEDEYGGFLNSGVINDFRDYTDLCFKEFGDRVRYWSTLNEPWVFSNSGYALGTNAPGR
FT                   CSASNVAKPGDSGTGPYIVTHNQILAHAEAVHVYKTKYQAYQKGKIGITLVSNWLMPLD
FT                   DNSIPDIKAAERSLDFQFGLFMEQLTTGDYSKSMRRIVKNRLPKFSKFESSLVNGSFDF
FT                   IGINYYSSSYISNAPSHGNAKPSYSTNPMTNISFEKHGIPLGPRAASIWIYVYPYMFIQ
FT                   EDFEIFCYILKINITILQFSITENGMNEFNDATLPVEEALLNTYRIDYYYRHLYYIRSA
FT                   IRAGSNVKGFYAWSFLDCNEWFAGFTVRFGLNFVD"
FT   mRNA            1..1859
FT                   /experiment="experimental evidence, no additional details
FT                   recorded"
XX
SQ   Sequence 1859 BP; 609 A; 314 C; 355 G; 581 T; 0 other;
     aaacaaacca aatatggatt ttattgtagc catatttgct ctgtttgtta ttagctcatt        60
     cacaattact tccacaaatg cagttgaagc ttctactctt cttgacatag gtaacctgag       120
     tcggagcagt tttcctcgtg gcttcatctt tggtgctgga tcttcagcat accaatttga       180
     aggtgcagta aacgaaggcg gtagaggacc aagtatttgg gataccttca cccataaata       240
     tccagaaaaa ataagggatg gaagcaatgc agacatcacg gttgaccaat atcaccgcta       300
     caaggaagat gttgggatta tgaaggatca aaatatggat tcgtatagat tctcaatctc       360
     ttggccaaga atactcccaa agggaaagtt gagcggaggc ataaatcacg aaggaatcaa       420
     atattacaac aaccttatca acgaactatt ggctaacggt atacaaccat ttgtaactct       480
     ttttcattgg gatcttcccc aagtcttaga agatgagtat ggtggtttct taaactccgg       540
     tgtaataaat gattttcgag actatacgga tctttgcttc aaggaatttg gagatagagt       600
     gaggtattgg agtactctaa atgagccatg ggtgtttagc aattctggat atgcactagg       660
     aacaaatgca ccaggtcgat gttcggcctc caacgtggcc aagcctggtg attctggaac       720
     aggaccttat atagttacac acaatcaaat tcttgctcat gcagaagctg tacatgtgta       780
     taagactaaa taccaggcat atcaaaaggg aaagataggc ataacgttgg tatctaactg       840
     gttaatgcca cttgatgata atagcatacc agatataaag gctgccgaga gatcacttga       900
     cttccaattt ggattgttta tggaacaatt aacaacagga gattattcta agagcatgcg       960
     gcgtatagtt aaaaaccgat tacctaagtt ctcaaaattc gaatcaagcc tagtgaatgg      1020
     ttcatttgat tttattggta taaactatta ctcttctagt tatattagca atgccccttc      1080
     acatggcaat gccaaaccca gttactcaac aaatcctatg accaatattt catttgaaaa      1140
     acatgggata cccttaggtc caaggggctgc ttcaatttgg atatatgttt atccatatat      1200
     gtttatccaa gaggacttcg agatcttttg ttacatatta aaaataaata taacaatcct      1260
     gcaattttca atcactgaaa atggtatgaa tgaattcaac gatgcaacac ttccagtaga      1320
     agaagctctt ttgaatactt acagaattga ttactattac cgtcacttat actacattcg      1380
     ttctgcaatc agggctggct caaatgtgaa gggtttttac gcatggtcat ttttggactg      1440
     taatgaatgg tttgcaggct ttactgttcg ttttggatta aactttgtag attagaaaga      1500
     tggattaaaa aggtacccta agctttctgc ccaatggtac aagaactttc tcaaaagaaa      1560
     ctagctagta ttattaaaag aactttgtag tagattacag tacatcgttt gaagttgagt      1620
     tggtgcacct aattaaataa aagaggttac tcttaacata tttttaggcc attcgttgtg      1680
     aagttgttag gctgttattt ctattatact atgttgtagt aataagtgca ttgttgtacc      1740
     agaagctatg atcataacta taggttgatc cttcatgtat cagtttgatg ttgagaatac      1800
     tttgaattaa aagtcttttt ttattttttt aaaaaaaaaa aaaaaaaaaa aaaaaaaaa       1859
//
```

```
ID   X56734; SV 1; linear; mRNA; STD; PLN; 1859 BP.
XX
AC   X56734; S46826;
XX
DT   12-SEP-1991 (Rel. 29, Created)
DT   25-NOV-2005 (Rel. 85, Last updated, Version 11)
XX
DE   Trifolium repens mRNA for non-cyanogenic beta-glucosidase
XX
KW   beta-glucosidase.
XX
OS   Trifolium repens (white clover)
OC   Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
OC   Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons; rosids;
OC   eurosids I; Fabales; Fabaceae; Papilionoideae; Trifolieae; Trifolium.
XX
RN   [5]
RP   1-1859
RX   DOI; 10.1007/BF00039495.
RX   PUBMED; 1907511.
RA   Oxtoby E., Dunn M.A., Pancoro A., Hughes M.A.;
RT   "Nucleotide and derived amino acid sequence of the cyanogenic
RT   beta-glucosidase (linamarase) from white clover (Trifolium repens L.)";
RL   Plant Mol. Biol. 17(2):209-219(1991).
XX
RN   [6]
RP   1-1859
RA   Hughes M.A.;
RT   ;
RL   Submitted (19-NOV-1990) to the EMBL/GenBank/DDBJ databases.
RL   Hughes M.A., University of Newcastle Upon Tyne, Medical School, Newcastle
RL   Upon Tyne, NE2 4HH, UK
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..1859
FT                   /organism="Trifolium repens"
FT                   /mol_type="mRNA"
FT                   /clone_lib="lambda gt10"
FT                   /clone="TRE361"
FT                   /tissue_type="leaves"
FT                   /db_xref="taxon:3899"
FT   CDS             14..1495
FT                   /product="beta-glucosidase"
FT                   /EC_number="3.2.1.21"
FT                   /note="non-cyanogenic"
FT                   /db_xref="GOA:P26204"
FT                   /db_xref="HSSP:1CBG"
FT                   /db_xref="InterPro:IPR018120"
FT                   /db_xref="UniProtKB/Swiss-Prot:P26204"
FT                   /protein_id="CAA40058.1"
FT                   /translation="MDFIVAIFALFVISSFTITSTNAVEASTLLDIGNLSRSSFPRGFI
FT                   FGAGSSAYQFEGAVNEGGRGPSIWDTFTHKYPEKIRDGSNADITVDQYHRYKEDVGIMK
FT                   DQNMDSYRFSISWPRILPKGKLSGGINHEGIKYYNNLINELLANGIQPFVTLFHWDLPQ
FT                   VLEDEYGGFLNSGVINDFRDYTDLCFKEFGDRVRYWSTLNEPWVFSNSGYALGTNAPGR
FT                   CSASNVAKPGDSGTGPYIVTHNQILAHAEAVHVYKTKYQAYQKGKIGITLVSNWLMPLD
FT                   DNSIPDIKAAERSLDFQFGLFMEQLTTGDYSKSMRRIVKNRLPKFSKFESSLVNGSFDF
FT                   IGINYYSSSYISNAPSHGNAKPSYSTNPMTNISFEKHGIPLGPRAASIWIYVYPYMFIQ
FT                   EDFEIFCYILKINITILQFSITENGMNEFNDATLPVEEALLNTYRIDYYYRHLYYIRSA
FT                   IRAGSNVKGFYAWSFLDCNEWFAGFTVRFGLNFVD"
FT   mRNA            1..1859
FT                   /experiment="experimental evidence, no additional details
FT                   recorded"
XX
```

```
SQ   Sequence 1859 BP; 609 A; 314 C; 355 G; 581 T; 0 other;
     aaacaaacca aatatggatt ttattgtagc catatttgct ctgtttgtta ttagctcatt        60
     cacaattact tccacaaatg cagttgaagc ttctactctt cttgacatag gtaacctgag       120
     tcggagcagt tttcctcgtg gcttcatctt tggtgctgga tcttcagcat accaatttga       180
     aggtgcagta aacgaaggcg gtagaggacc aagtatttgg gataccttca cccataaata       240
     tccagaaaaa ataagggatg gaagcaatgc agacatcacg gttgaccaat atcaccgcta       300
     caaggaagat gttgggatta tgaaggatca aaatatggat tcgtatagat tctcaatctc       360
     ttggccaaga atactcccaa agggaaagtt gagcggaggc ataaatcacg aaggaatcaa       420
     atattacaac aaccttatca acgaactatt ggctaacggt atacaaccat ttgtaactct       480
     ttttcattgg gatcttcccc aagtcttaga agatgagtat ggtggtttct taaactccgg       540
     tgtaataaat gattttcgag actatacgga tctttgcttc aaggaatttg gagatagagt       600
     gaggtattgg agtactctaa atgagccatg ggtgtttagc aattctggat atgcactagg       660
     aacaaatgca ccaggtcgat gttcggcctc caacgtggcc aagcctggtg attctggaac       720
     aggaccttat atagttacac acaatcaaat tcttgctcat gcagaagctg tacatgtgta       780
     taagactaaa taccaggcat atcaaaaggg aaagataggc ataacgttgg tatctaactg       840
     gttaatgcca cttgatgata atagcatacc agatataaag gctgccgaga gatcacttga       900
     cttccaattt ggattgttta tggaacaatt aacaacagga gattattcta agagcatgcg       960
     gcgtatagtt aaaaaccgat tacctaagtt ctcaaaattc gaatcaagcc tagtgaatgg      1020
     ttcatttgat tttattggta taaactatta ctcttctagt tatattagca atgccccttc      1080
     acatggcaat gccaaaccca gttactcaac aaatcctatg accaatattt catttgaaaa      1140
     acatgggata cccttaggtc caagggctgc ttcaatttgg atatatgttt atccatatat      1200
     gtttatccaa gaggacttcg agatcttttg ttacatatta aaaataaata taacaatcct      1260
     gcaattttca atcactgaaa atggtatgaa tgaattcaac gatgcaacac ttccagtaga      1320
     agaagctctt ttgaatactt acagaattga ttactattac cgtcacttat actacattcg      1380
     ttctgcaatc agggctggct caaatgtgaa gggtttttac gcatggtcat ttttggactg      1440
     taatgaatgg tttgcaggct ttactgttcg ttttggatta aactttgtag attagaaaga      1500
     tggattaaaa aggtacccta agctttctgc ccaatggtac aagaactttc tcaaaagaaa      1560
     ctagctagta ttattaaaag aactttgtag tagattacag tacatcgttt gaagttgagt      1620
     tggtgcacct aattaaataa aagaggttac tcttaacata tttttaggcc attcgttgtg      1680
     aagttgttag gctgttattt ctattatact atgttgtagt aataagtgca ttgttgtacc      1740
     agaagctatg atcataacta taggttgatc cttcatgtat cagtttgatg ttgagaatac      1800
     tttgaattaa aagtcttttt ttatttttt aaaaaaaaaa aaaaaaaaaa aaaaaaaaa        1859
//
```

International DNA databases

Genbank at NCBI

http://www.ncbi.nlm.nih.gov/genbank/

ENA at EMBL-EBI

http://www.ebi.ac.uk/embl/ena

DDBJ in Japan

http://www.ddbj.nig.ac.jp/

International Nucleotide Sequence Database Collaboration

http://www.insdc.org/

DATA sources for DNA databases

- Direct scientist submission
- Genome sequencing labs and groups
- Scientific literature
- Patent applications

- GenBank, ENA and DDBJ collaborate to collect all sequence data reported around the world.

International DNA databases

All of these databases have:

Official releases every 2-3 months.

Weekly (or daily updates).

Are divided into sublibraries for easier searching.

DNA database divisions

- PRI - primate (human,monkey)
- ROD - rodent (mouse,rat)
- MAM - other mammalian (bovine,cat)
- VRT - other vertebrate (chicken)
- INV - invertebrate
- PLN - plant, fungal, and alga
- BCT - bacteria
- VRL - viruses
- PHG - bacteriophage
- SYN - synthetic (plasmids, vectors)
- UNA - unannotated sequences
- PAT - patent sequences

- EST - Expressed Sequence Tags
- STS - Sequence Tagged Sites
- GSS - Genome Survey Sequences
- HTG - High Throughput Genomic Sequences
- HTC - High Throughput cDNA Sequences

Short Read and Trace Archives

The output of large scale sequencing projects and next-generation sequencing are stored in separate databases.

The Trace Archive is for Sanger reads
SRA is for next-generation sequencing techonologies

Genomic databases

- Specialized resources that are:
  - Species specific
  - Sequencing technique specific
- Display whole chromosomes (not a specific sequence).

RefSeq: NCBI Reference Sequence Database

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

The Reference Sequence (RefSeq) collection aims to provide a comprehensive, integrated, non-redundant set of sequences, including genomic DNA, transcript (RNA), and protein products, for major research organisms

https://www.ncbi.nlm.nih.gov/refseq/

REFSEQ from NCBI

- non-redundancy
- explicitly linked nucleotide and protein sequences
- updates to reflect current knowledge of sequence data and biology
- data validation and format consistency
- distinct accession series
- ongoing curation by NCBI staff and collaborators, with reviewed records indicated

RefSeq record Status

- The RefSeq COMMENT block indicates the Status of the record and the GenBank sequence data that was used to provide the record.

- In addition, the COMMENT may identify a collaboration which supplied the defining sequence information for the genome, gene, or protein.

  The level of curation may differ between different collaborating groups.

RefSeq

✳Status Codes:

RefSeq records are provided with a status code which provides an indication of the level of review a RefSeq record has undergone.

- Reviewed*
- Provisional
- Predicted
- Genome Annotation
- Validated*
- Model
- Inferred
- WGS

* Curated

STATUS	Definition
REVIEWED	The RefSeq record has been the reviewed by NCBI staff or by collaborator. The NCBI review process includes reviewing available sequence data and frequently also includes a review of the literature and other sources of information.
VALIDATED	The RefSeq record has undergone an initial review to provide the preferred sequence standard. The record has not yet been subject to final review at which time additional functional information may be provided.
PROVISIONAL	The RefSeq record has not yet been subject to individual review and is thought to be well supported and to represent a valid transcript and protein.

STATUS	Definition
PREDICTED	The RefSeq record is predicted and has not been subject to individual review. The transcript may represent an *ab initio* prediction or may be partially supported by other transcript data; in both cases, the protein is predicted.
INFERRED	The RefSeq record is inferred by genome sequence analysis. There is no same-organism experimental support for the full extent of the sequence; there may be some level of support by homology.
MODEL	The RefSeq record is predicted by genome sequence analysis. The record may represent an *ab initio* prediction, or may have some level of transcript or homology support.

STATUS	Definition
GENOME ANNOTATION	This identifies RefSeq records provided by the NCBI Genome Annotation process. These records are provided via automated processing and are not subject to individual review or revision between builds
WGS	The RefSeq record represents a collection of whole genome shotgun (WGS) sequences. This status code is applied to genomic records

Accession Format	Molecule Type
NC_123456	Complete Genome Complete Chromosome Complete Sequence
NG_123456	Genomic Region
NM_123456	mRNA
NR_123456	non-coding RNA
NP_123456	Protein
NT_123456	Genomic Contig (from BACs)
NW_123456	Genomic Contig (from WGS)
XM_123456	mRNA (taken from genomic seq)
XR_123456	RNA (taken from genomic seq)
XP_123456	Protein (taken from genomic seq)

What is the difference between RefSeq and GenBank?

Genbank :

- Archival database and includes publicly available DNA sequences submitted from individual laboratories and large-scale sequencing projects.
- Submitted sequence data is exchanged between NCBIs GenBank, EMBL Data Library (EMBL) and the DNA Data Bank of Japan (DDBJ) to achieve comprehensive worldwide coverage.
- As an archival database, GenBank is very redundant for some loci.
- Sequence records are owned by the original submitter and can not be altered by a third party.

What is the difference between RefSeq and GenBank?

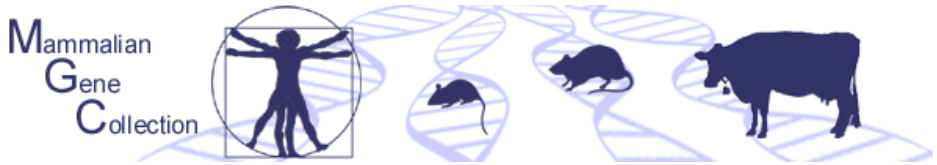
RefSeq :

- Sequences are derived from GenBank and provide non-redundant curated data.
- Entries records represent current knowledge.
- RefSeq records are owned by NCBI and therefore can be updated as needed to maintain current annotation or to incorporate additional sequence information.
- Some sequence records are provided through collaboration; and thus may not be available in any one GenBank record (not primary)

Various High Throughput Collections

Nedo, DFKZ, HRI, Genoscope

- Full-length cDNA libraries from various tissues were subtracted and normalized to reduce redundancy
- Clones were end-sequenced to further reduce redundancy
- Whole inserts were sequenced to get mRNA sequences
- [KIAA – done by Kazusa was a project for long cDNAs – over 4kb, but may not be full-length]



MGC - Mammalian Gene Collection

The NIH Mammalian Gene Collection (MGC) seeks to identify and sequence a representative full open reading frame (ORF) clone for each human, mouse, rat and cow gene. Zebrafish and Xenopus have their own projects (ZGC and XGC)

MGC produced over 80 cDNA libraries enriched for full-length cDNAs derived from human tissue and cell lines, and mouse tissue.

5' EST reads were generated from each library. Several algorithms are applied to select putative full ORF clones. Targeted cloning or synthesis was used to finish.

Accession Numbers

	Accession Numbers
• Individual Labs	various
• Refseq	XX_123456

Full Length Sequencing projects:

• Riken, Nedo (FLJ), HRI	AK, CR
DKFZ, Genoscope, [KIAA]...	[AB, D]
• MGC	BC, CT

Gene symbols

Gene symbols are designated by upper case Latin letters or by a combination of upper-case letters and Arabic numbers.

Symbols should be short in order to be useful, and should not attempt to represent all known information about a gene.

Based on classical genetic guidelines, it is recommended that gene symbols are either underlined or italicized when referring to genotypic information (phenotypic information is represented in standard fonts).

HUGO Gene Nomenclature Committee

- This committee is responsible for the approval of a unique symbol for each gene.
- It also designs a longer and more descriptive name.

- The committee makes considerable efforts to use symbols acceptable to workers in the field, but sometimes it is not possible to use exactly what has previously appeared in the literature.
- However, wherever the committee is aware of such symbols, they are listed as aliases

Taxonomy Databases

- An international effort is done for all sequence databases to create a unified taxonomic tag for the sequences submitted.

- Problem: each sequence depositor gives "his" name for the species
- Solution: Unified taxonomy ID

Gene

Gene integrates information from a wide range of species. A record may include nomenclature, Reference Sequences (RefSeqs), maps, pathways, variations, phenotypes, and links to genome-, phenotype-, and locus-specific resources worldwide.

With the increasing sequencing and annotation of key genomes, having a gene-based view of the resultant information is useful. Entrez Gene has therefore been implemented to supply key connections in the nexus of map, sequence, expression, structure, function, citation, and homology data. Unique identifiers are assigned to genes with defining sequences, genes with known map positions, and genes inferred from phenotypic information.

Entrez Gene at NCBI

Entrez Gene - A database for gene-specific information.

It does not include all known or predicted genes; instead **Entrez Gene** focuses on the genomes that have been completely sequenced, that have an active research community to contribute gene-specific information, or that are scheduled for intense sequence analysis.

The content of Entrez Gene represents the result of curation and automated integration of data from NCBI's Reference Sequence project (RefSeq), from collaborating model organism databases, and from many other databases available from NCBI. Records are assigned unique, stable and tracked integers as identifiers.

Entrez Gene at NCBI

The content (nomenclature, map location, gene products and their attributes, markers, phenotypes, and links to citations, sequences, variation details, maps, expression, homologs, protein domains and external databases) is updated as new information becomes available.

Entrez Gene data is used by other NCBI resources such as: BLAST, Geo, HomoloGene, Map Viewer, UniGene, UniSTS and NCBI's genome annotation pipeline.

Protein databases

Protein databases

- There are many different protein databases containing different types of information:
  - Primary Amino Acids sequence.
  - Secondary structure
  - 3D structure
  - Protein family domains
  - Consensus active sites
- Usually contain description of the protein entry (annotation), the amino acid sequence and sometimes links to other related databases.

Sources of Protein

- Proteins that have been worked on experimentally
- mRNA whose product has been worked on experimentally (no actual protein sequencing done)
- Translated DNA (mRNA) sequences

UniProt: Universal Protein Resource

UniProt is the most comprehensive catalog of information on proteins. It is a central repository of protein sequence and function created by joining the information contained in Swiss-Prot, TrEMBL, and PIR.

The UniProt Knowledgebase (UniProt) is the central access point for extensive curated protein information, including function, classification, and cross-reference.

Swiss-Prot Database (primary database)

- Swiss-Prot annotation includes:
  - Description of protein function
  - Subcellular localization
  - Protein domain structure
  - Post-translational modifications
  - Protein variants

Swiss-Prot Database

Swiss-Prot differs from other protein databases by the following criteria:

- Annotation
- Minimal Redundancy
- Integration with other databases

The annotation consists of the description of:

- Function(s) of the protein
- Post-translational modification(s). For example carbohydrates, phosphorylation, acetylation, GPI-anchor, etc.
- Domains and sites. For example calcium binding regions, ATP-binding sites, zinc fingers, etc.
- Secondary structure

The annotation consists of the description of:

- Quaternary structure. For example homodimer, heterotrimer, etc.
- Similarities to other proteins
- Disease(s) associated with deficiency(s) of/in the protein
- Sequence conflicts, variants, etc.

Swiss-Prot Database

To obtain this information, Swiss-Prot uses, in addition to the publications that report new sequence data, review articles to periodically update the annotations of families or groups of proteins.

Swiss-Prot also makes use of external experts, who have been recruited to send their comments and updates concerning specific groups of proteins.

Swiss-Prot Database

- Minimal Redundancy

  Many sequence databases contain, for a given protein sequence, separate entries which correspond to different literature reports.
  In SWISS-PROT, they try as much as possible to merge all these data so as to minimize the redundancy of the database.

  If conflicts exist between various sequencing reports, they are indicated in the feature table of the corresponding entry.

Swiss-Prot Database

- Integration with other databases

  It is important to provide the users of biomolecular databases with a degree of integration between the three types sequence-related databases (nucleic acid sequences, protein sequences and protein tertiary structures) as well as with specialized data collections.

  SWISS- PROT is currently cross-referenced with ~100 different databases. Cross-references are provided in the form of pointers to information related to SWISS-PROT entries and found in data collections other than SWISS-PROT.

TrEMBL database

- TrEMBL is a computer-annotated supplement of SWISS-PROT that contains all the translations of the EMBL (DNA) database.

- TrEMBL contain entries not yet integrated in SWISS-PROT.

Data reliability in databases

The huge amount of data collected in databases present a lot of problems:

- Data accuracy
- Sequence redundancy
- Inconsistent nomenclature
- Inaccurate annotation
- Sequence contamination (vectors, bacterial)

Data reliability in databases

- The database staff notify the Authors that an error (or contamination) was detected in their sequence entry.
- However, it takes time to correct the data.
- Meanwhile the error is continued, because annotation is transferred from database to database.

Data reliability in databases

- A lot of the sequences in the database are quite “old”. They were not updated since they were submitted, even though technology and data was very much updated.

Data reliability in Protein databases

- Many proteins in the databases have erroneous sequences due to:
  - missing exons in the DNA translation.
  - Introns mistakenly translated.

- Another common problem is the assigning of functions to “new” proteins, based on sequence similarity.

Data reliability in Protein databases

- For example:
  - Protein A is similar to protein B.
  - Protein B annotation is based on Protein A annotation (which has an error).
  - Annotation of Protein A is corrected by the group working on it. This correction does not appear or reflect in Protein B annotation.
  - When Protein C and D are also based on the erroneous annotation on B, the problem……

Text searching pitfalls

- It finds exactly what you type (try pseudogene vs. psuedogene)
- Older records may have different annotation, from gene names on…
- human vs homo sapiens
- Gene symbols vs full gene name (for example neuregulin vs nrg1)

- Most sites use boolean operators (AND, OR, BUT NOT)
- Can do (or add) a field specific tag - but each site has a different way of adding it to a search - for example, NCBI uses square brackets []

Remember:

Text searching is NOT sequence similarity searching! You many not find all related sequences by text searching!!!!