LSM2241
Introduction to Bioinformatics Databases and Tools

Greg Tucker-Kellogg

12 August 2015

Outline

What are databases?

Finding out about databases

DNA sequence databases and GenBank

Reference databases

Databases and their limits

Round up and next week

Topic

What are databases?

Finding out about databases

DNA sequence databases and GenBank

Reference databases

Databases and their limits

Round up and next week

Key database concepts

Entity	The ”thing” (perhaps something in the real world) being represented as data in a database
Record	A collection of related data about an entity, kept and stored together. A record is usually arranged as a collection of *fields*
Field	Part of a record used for a particular type of data
Relation	Structured data link between different entities (or records)

An example ”grocery shopping” database

Grocery Database

	A	B	C	D
1	Date	Item type	Store	Price
2				
3	18/08/2006	White bread	Dover Provision	$1.29
4	18-Aug-06	Eggs	Dover	$1.49
5	August 18 '06	Milk	Minijumbo Provision	2.49
6	29/07/2006	Bread	Cold Storage	$1.39
7				

What if you need *huge* number of fields for a record type?

- Most of the fields of a record would be empty!
- An Entity-Attribute-Value model can be used instead (example from Wikipedia). This only requires three fields!

```
(<patient XYZ, 1/5/98 9:30 AM>, <Temperature in degrees Fahrenheit>, "102")
(<patient XYZ, 1/5/98 9:30 AM>, <Presence of Cough>, "True" )
(<patient XYZ, 1/5/98 9:30 AM>, <Type of Cough>, "With phlegm, yellowish, streaks o
(<patient XYZ, 1/5/98 9:30 AM>, <Heart Rate in beats per minute>, "98" )
```

Some features of biological databases

- Stored in records according to some predetermined structure or format
  - +/- evidence
  - +/- unique identifiers
  - +/- additional annotation
  - +/- database cross references
- Most of them are text data in “raw” form
- Strings, numbers, and dates
- “Strings” and sequences mean a set of consecutive characters
  - GCGCAATTCGCG
  - AKT is a protein kinase
  - 月亮代表我的心

Example: p53 *gene* record from GenBank

TP53 tumor protein p53 [*Homo sapiens*]

Gene ID: 7157, updated on 13-Jan-2013

Summary

Official Symbol	TP53 provided by HGNC
Official Full Name	tumor protein p53 provided by HGNC
Primary source	HGNC:11998
See related	Ensembl:ENSG00000141510; HPRD:01859; MIM:191170; Vega:OTTHUMG00000162125
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Homo sapiens
Lineage	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as	P53; BCC7; LFS1; TRP53
Summary	This gene encodes tumor protein p53, which responds to diverse cellular stresses to regulate target genes that induce cell cycle arrest, apoptosis, senescence, DNA repair, or changes in metabolism. p53 protein is expressed at low level in normal cells and at a high level in a variety of transformed cell lines, where it's believed to contribute to transformation and malignancy. p53 is a DNA-binding protein containing transcription activation, DNA-binding,

The human p53 gene record

Topic

What are databases?

Finding out about databases

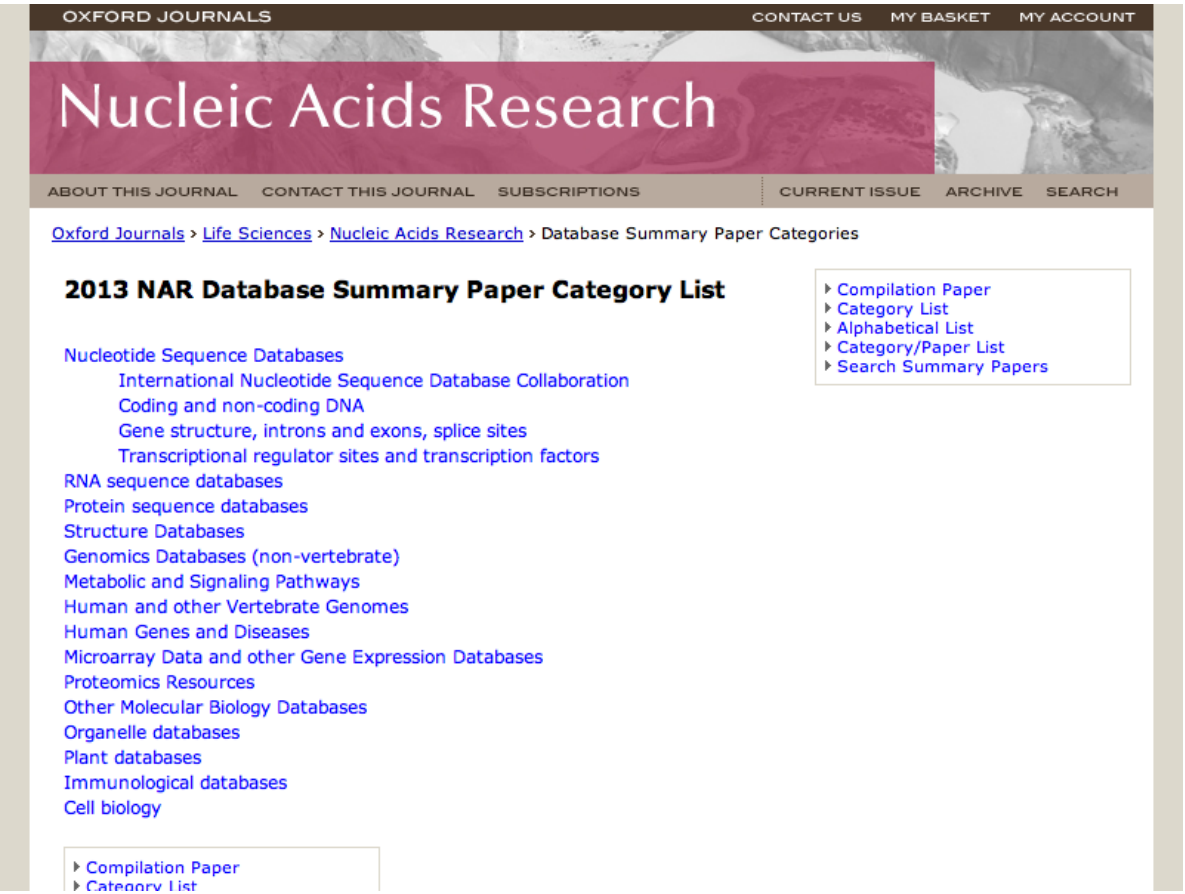
DNA sequence databases and GenBank

Reference databases

Databases and their limits

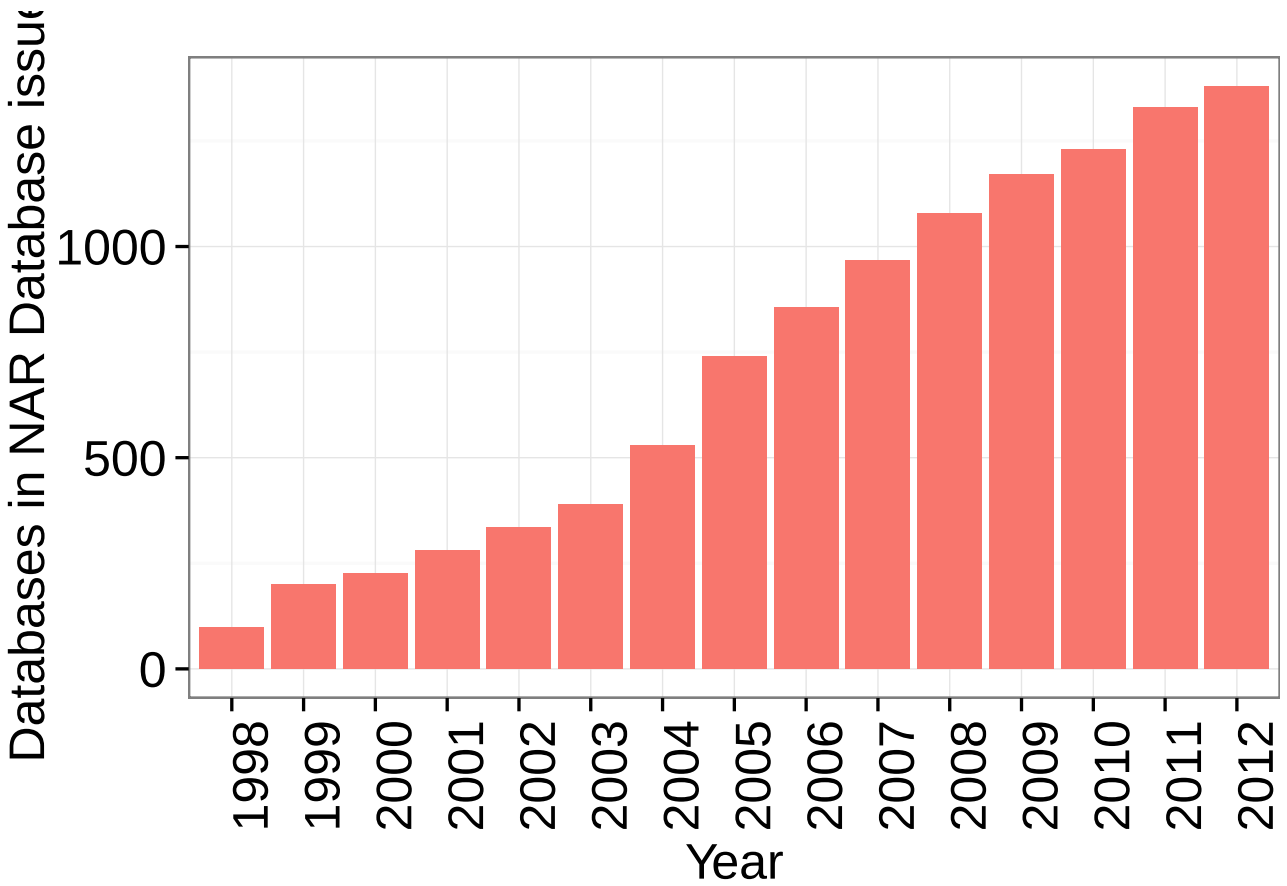
Round up and next week

The Nucleic Acids Research annual database issue



The NAR database issue

How many biological databases are there?



Growth in NAR database issue listed databases

The NCBI Entrez system

- The US National Center for Biotechnology Information (NCBI) hosts an essential collection of web-based resources
- NCBI Entrez searches many of these at once
- We will use these early and often throughout this module
- The first one we will use is PubMed, starting this week
- Anyone can use the NCBI resources.
- To use them most effectively, you should register an NCBI account (can use Google login for this)
- There is an excellent PubMed tutorial available
- The NCBI has a YouTube channel filled with video tutorials

A reference management tool

- The product of science is published in articles
- You need to cite articles and take effective notes
- You need to create bibliographies
- Reference management software is designed to do this
- Mendeley, EndNote, and WizFolio are three tools to do this
- We will maintain a Mendeley group of references for every reference used in this module

Topic

What are databases?

Finding out about databases

DNA sequence databases and GenBank

Reference databases

Databases and their limits

Round up and next week

Primary and secondary databases

Primary	experimental results, possibly with curation, review, or error correction
Secondary	Databases containing data derived from experimental data

The big three sequence databases

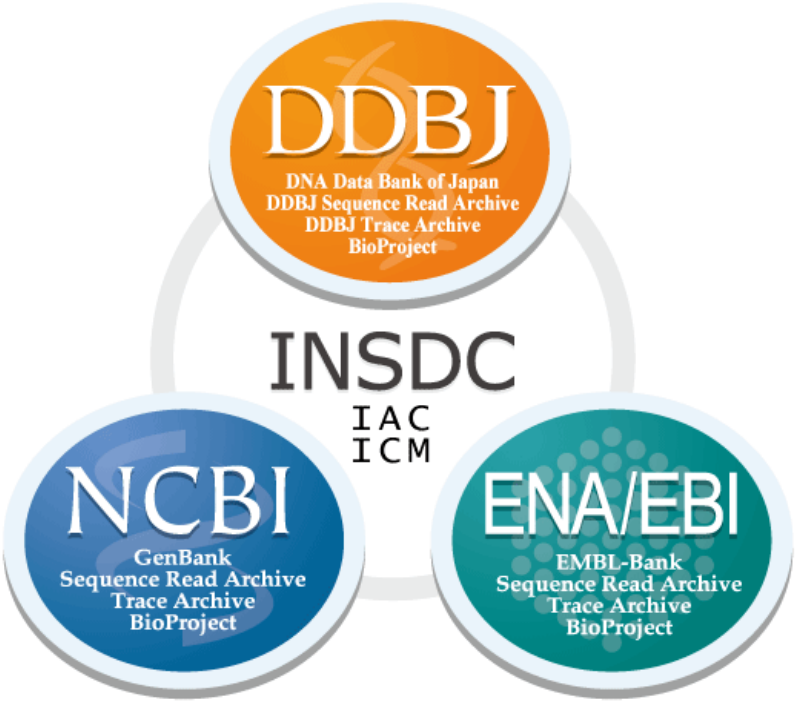
Three major databases coordinate and share sequence information

1. The National Centre for Biotechnology Information (USA, NCBI)
2. The DNA Data Bank of Japan (DDBJ),
3. The European Nucleotide Archive at theEMBL-EBI (ENA)

They share information daily

But there are many more!

The International Nucleotide Sequence Database Collaboration



The Big Three

Accession numbers

- Database information can change, but sequence data is important!
- All the major sequence databases define *Accession numbers* that can be used as a unique, relatively stable, identifier of a biological sequence
- Genes have *symbols*, sequences have *accession numbers*
- Accession numbers have *versions*
- An accession number + version is guaranteed to be an unchanging sequence

Genbank

- A database of most known public sequences (DNA and protein)
  - more than 300,000 organisms
  - more than 161,000,000 sequences
- Each record includes
  - The sequence itself
  - Biological annotation
  - Literature information

Let's take a look at human p53

- Which sequence should we use? There are several transcripts and protein variants
- How did we get here?

Homo sapiens tumor protein p53 (TP53), transcript variant 1, mRNA

NCBI Reference Sequence: NM_000546.5

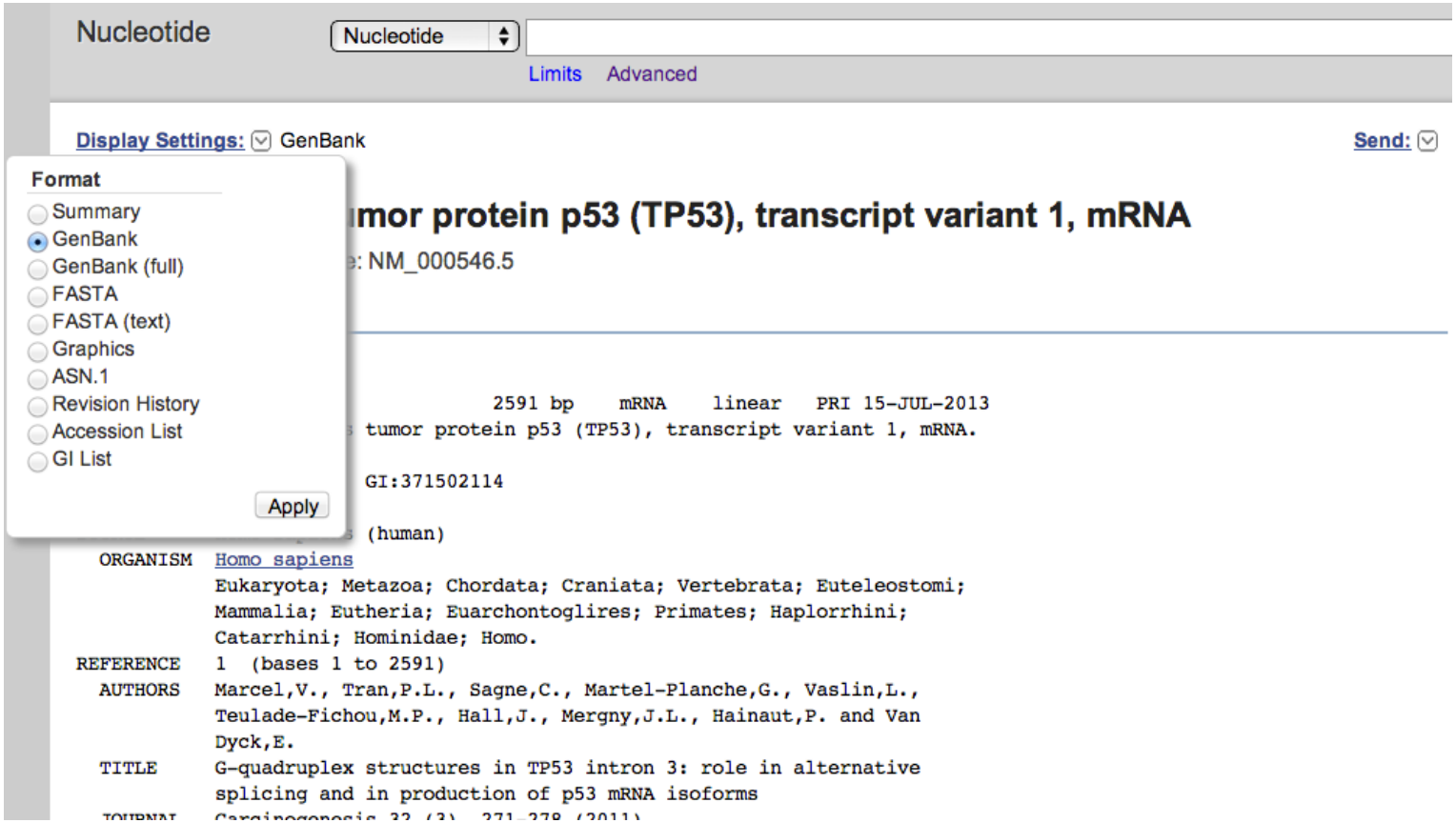
FASTA Graphics

Go to:

```
LOCUS       NM_000546               2591 bp    mRNA    linear   PRI 23-DEC-2012
DEFINITION  Homo sapiens tumor protein p53 (TP53), transcript variant 1, mRNA.
ACCESSION   NM_000546
VERSION     NM_000546.5  GI:371502114
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 2591)
  AUTHORS   Marcel,V., Tran,P.L., Sagne,C., Martel-Planche,G., Vaslin,L.,
            Teulade-Fichou,M.P., Hall,J., Mergny,J.L., Hainaut,P. and Van
            Dyck,E.
  TITLE     G-quadruplex structures in TP53 intron 3: role in alternative
            splicing and in production of p53 mRNA isoforms
  JOURNAL   Carcinogenesis 32 (3), 271-278 (2011)
```

Human p53 Transcript variant 1

Display settings affect what you see



If you want to download a FASTA formatted sequence, use FASTA (text)

Identifying information

Homo sapiens tumor protein p53 (TP53), transcript variant 1, mRNA

NCBI Reference Sequence: NM_000546.5

FASTA Graphics

Go to:

```
LOCUS       NM_000546                2591 bp    mRNA    linear   PRI 23-DEC-2012
DEFINITION  Homo sapiens tumor protein p53 (TP53), transcript variant 1, mRNA.
ACCESSION   NM_000546
VERSION     NM_000546.5  GI:371502114
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
```

Identifying information

The header information of a GenBank record

Sources of information

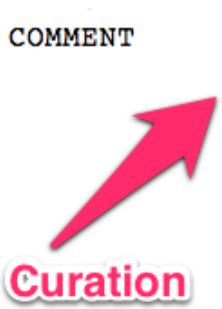
```
ACCESSION   NM_000546
VERSION     NM_000546.5  GI:371502114
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 2591)
  AUTHORS   Marcel,V., Tran,P.L., Sagne,C., Martel-Planche,G., Vaslin,L.,
            Teulade-Fichou,M.P., Hall,J., Mergny,J.L., Hainaut,P. and Van
            Dyck,E.
  TITLE     G-quadruplex structures in TP53 intron 3: role in alternative
            splicing and in production of p53 mRNA isoforms
  JOURNAL   Carcinogenesis 32 (3), 271-278 (2011)
   PUBMED   21112961
```

organism and taxonomy

Literature

Just below the header fields

The ever important COMMENT field




```
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence was derived from DA453049.1, X02469.1,
            AK223026.1 and DQ186650.1.
            This sequence is a reference standard in the RefSeqGene project.
            On Jan 7, 2012 this sequence version replaced gi:187830767.

            Summary: This gene encodes tumor protein p53, which responds to
            diverse cellular stresses to regulate target genes that induce cell
            cycle arrest, apoptosis, senescence, DNA repair, or changes in
            metabolism. p53 protein is expressed at low level in normal cells
            and at a high level in a variety of transformed cell lines, where
            it's believed to contribute to transformation and malignancy. p53
            is a DNA-binding protein containing transcription activation,
            DNA-binding, and oligomerization domains. It is postulated to bind
            to a p53-binding site and activate expression of downstream genes
            that inhibit growth and/or invasion, and thus function as a tumor
```

Sequences may be composed of other primary sequences

```
PRIMARY      REFSEQ_SPAN         PRIMARY_IDENTIFIER PRIMARY_SPAN        COMP
             1-67                DA453049.1         1-67
             68-1384             X02469.1           1-1317
             1385-2572           AK223026.1         478-1665
             2573-2591           DQ186650.1         2257-2275
FEATURES             Location/Qualifiers
     source          1..2591
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
```

The primary sequences of the p53 primary transcript

Features

```
FEATURES             Location/Qualifiers
     source          1..2591
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
                     /chromosome="17"
                     /map="17p13.1"
     gene            1..2591
                     /gene="TP53"
                     /gene_synonym="BCC7; LFS1; P53; TRP53"
                     /note="tumor protein p53"
                     /db_xref="GeneID:7157"
                     /db_xref="HGNC:11998"
                     /db_xref="MIM:191170"
```

many features may appear in one sequence (Do you remember what is this called?)

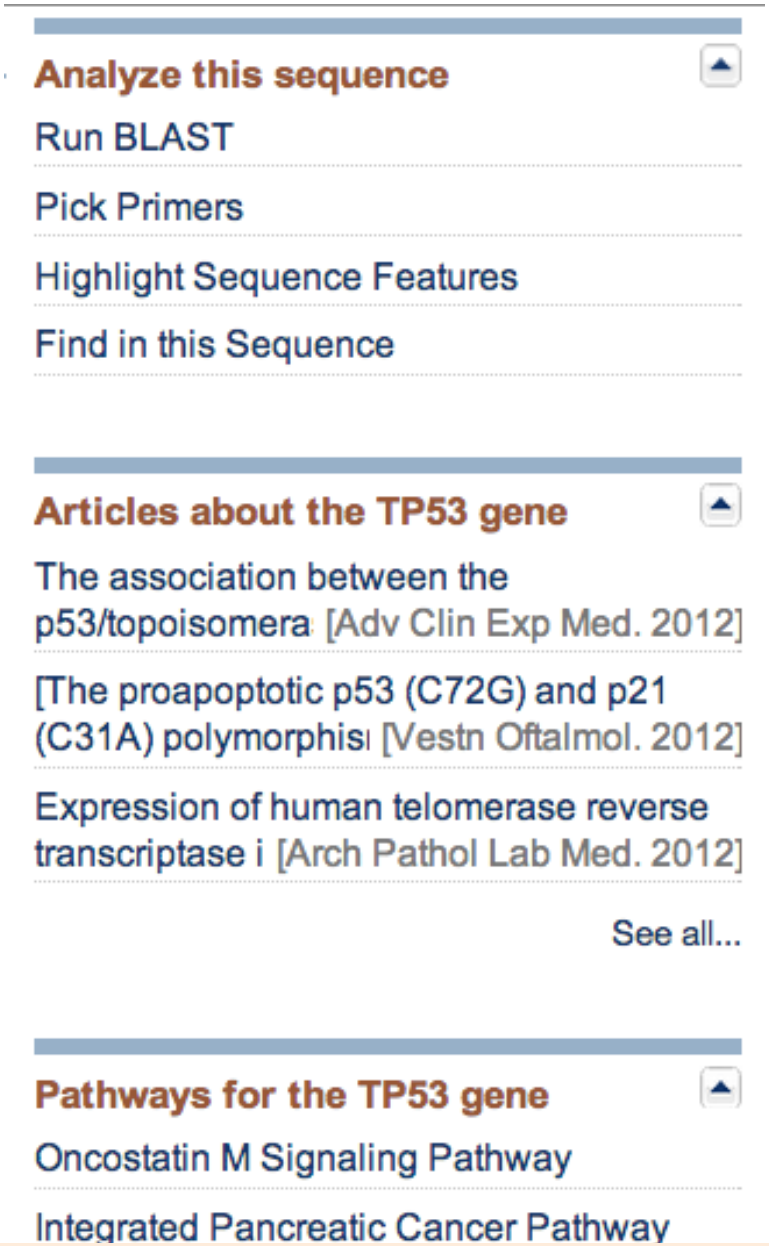
Sequence

```
ORIGIN
        1 gatgggattg gggttttccc ctcccatgtg ctcaagactg gcgctaaaag ttttgagctt
       61 ctcaaaagtc tagagccacc gtccagggag caggtagctg ctgggctccg gggacacttt
      121 gcgttcgggc tgggagcgtg ctttccacga cggtgacacg cttccctgga ttggcagcca
      181 gactgccttc cgggtcactg ccatggagga gccgcagtca gatcctagcg tcgagccccc
      241 tctgagtcag gaaacatttt cagacctatg gaaactactt cctgaaaaca acgttctgtc
      301 ccccttgccg tcccaagcaa tggatgattt gatgctgtcc ccggacgata ttgaacaatg
      361 gttcactgaa gacccaggtc cagatgaagc tcccagaatg ccagaggctg ctccccccgt
      421 ggcccctgca ccagcagctc ctacaccggc ggcccctgca ccagccccct cctggcccct
      481 gtcatcttct gtcccttccc agaaaaccta ccagggcagc tacggtttcc gtctgggctt
      541 cttgcattct gggacagcca agtctgtgac ttgcacgtac tcccctgccc tcaacaagat
      601 gttttgccaa ctggccaaga cctgccctgt gcagctgtgg gttgattcca cacccccgcc
      661 cggcacccgc gtccgcgcca tggccatcta caagcagtca cagcacatga cggaggttgt
```

The sequence itself is at the END of an entry

Links

- The links available through Entrez show up on the right
- *Analysis* options also appear there
- Different entries may have vastly differing numbers of links
- Links between protein and transcript records are very useful



Different kinds of nucleic acid sequence data

- Genomic DNA
- RNA (written as DNA sequence in the database!)
- Expressed sequence tags (ESTs)
  - single pass sequencing of cDNA clones
- UniGene entries are clusters of ESTs to identify non-redundant expressed genes
- Sequence tagged sites (STS)

Topic

What are databases?

Finding out about databases

DNA sequence databases and GenBank

Reference databases

Databases and their limits

Round up and next week

Getting from the transcript to the protein in Genbank

cellular tumor antigen p53 isoform a [Homo sapiens]

NCBI Reference Sequence: NP_000537.3

FASTA Graphics

Go to:

```
LOCUS       NP_000537                393 aa            linear   PRI 23-DEC-2012
DEFINITION  cellular tumor antigen p53 isoform a [Homo sapiens].
ACCESSION   NP_000537
VERSION     NP_000537.3  GI:120407068
DBSOURCE    REFSEQ: accession NM_000546.5
KEYWORDS    .
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (residues 1 to 393)
  AUTHORS   Marcel,V., Tran,P.L., Sagne,C., Martel-Planche,G., Vaslin,L.,
            Teulade-Fichou,M.P., Hall,J., Mergny,J.L., Hainaut,P. and Van
            Dyck,E.
```

How do we know this is the right protein?

With so many records how do we know which one to work with?

- They may come from different source databases (e.g., DDBJ, GenBank, EMBL)
- They may have the same or different sequence information
- They may have single changes in nucleotides/amino acids
- They may have variable extra annotatio (e.g., signal peptide; domains; cross-references, etc.)

The RefSeq Project

Goal

A ”comprehensive, integrated, non-redundant set of sequences, including genomic DNA, transcript (RNA), and protein products, for major research organisms.”
http://www.ncbi.nlm.nih.gov/RefSeq/index.html

Information from

- Predictions from genomic sequence
- Analysis of GenBank Records
- Collaborating databases

In what way are RefSeq entries different from Genbank entries

- non-redundant
- stable
- Curated
- Systematic (genomic DNA, transcripts, proteins)

Interpreting RefSeq identifiers

The first two letters of a RefSeq Accession tell its data type

Prefix	meaning
NC	Complete genomic molecule, usually reference assembly
NG	Incomplete genomic region
NM	mRNA
NP	Protein
NT	contig or scaffold sequence

Interpreting RefSeq status codes

Code	Description
VALIDATED	Has undergone an initial review to provide the preferred sequence standard. The record has not yet been subject to final review
REVIEWED	Has been reviewed by NCBI staff or by a collaborator. The NCBI review process includes assessing available sequence data and the literature.
PROVISIONAL	Has not yet been subject to individual review. The initial sequence-to-gene association has been established by outside collaborators or NCBI staff
PREDICTED	Has not yet been subject to individual review, and some aspect of the RefSeq record is predicted
MODEL	Is provided by the NCBI Genome Annotation pipeline and is not subject to individual review or revision between annotation runs
INFERRED	Has been predicted by genome sequence analysis, but it is not yet supported by experimental evidence. The record may be partially supported by homology data.
WGS	Is provided to represent a collection of whole genome shotgun sequences. These records are not subject to individual review or revisions between genome updates.

SwissProt and TrEMBL

- Swiss-Prot was established as a manually curated and updated database of protein sequences
  - Trained biologists extract and analyze relevant evidence from scientific publications
  - Include post-translational modifications, sequence variations, functions, etc
- TrEMBL is translated EMBL sequence entries
- UniProtKB is SwissProt + TrEMBL

The UniprotKB Knowledge Base of Proteins

UniProt

Downloads · Contact · Documentation/Help

Search | Blast | Align | Retrieve | ID Mapping

Search in: Protein Knowledgebase (UniProtKB)

Query: Search | Advanced Search » | Clear

THE FIRST 10 YEARS OF UNIPROT

Saturday 8 September 2012, Basel, Switzerland

Celebrate UniProt's 10th anniversary with us! Renowned speakers from the fields of interactions & protein modelling, proteomics, protein structure & function, and genome analysis & annotation will highlight how protein databases are underpinning life sciences.

Submit abstracts for talks and posters to abstractsUP12@isb-sib.ch

WELCOME

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

What we provide

UniProtKB	Protein knowledgebase, consists of two sections: Swiss-Prot, which is manually annotated and reviewed. TrEMBL, which is automatically annotated and is not reviewed

NEWS

UniProt release 2012_07 - Jul 11, 2012

To pee or not to pee I Removal of the cross-reference to CMR

- Statistics for UniProtKB: Swiss-Prot · TrEMBL
- Forthcoming changes
- News archives

Follow @uniprot 378 followers

SITE TOUR

Learn how to make best use of the tools and data on this site.

UNIPROT is a catalog of information on proteins

Topic

What are databases?

Finding out about databases

DNA sequence databases and GenBank

Reference databases

Databases and their limits

Round up and next week

What next?

- Interaction databases
- Pathway databases
- Disease databases
- Gene expression databases

Limitations of biological databases (case by case)

- redundancy
- incompletness
- Errors
- *Derived* data versus primary data
- Active update
  - Including new data
  - Including corrections to old data
  - Including proper links
- Many databases in older issues of NAR database issue are no longer updated

Database record retractions

- Our database entry for p53 primary transcript 1 was `NM_000546.5`. This means that versions 0-4 are now obsolete!
- Versions 1-4 are retracted
- If you record the accession number without the version, it may change

Topic

What are databases?

Finding out about databases

DNA sequence databases and GenBank

Reference databases

Databases and their limits

Round up and next week

What we've learned

- With explosive growth in sequence data, the major global sequence databases in the USA, Europe, and Japan provide shared information to public data.
- Deciding what fields become part of a database record is an ongoing challenge
- The RefSeq project assigns both sequence type and quality assessment to curated, non-redundant sequences
- Accession numbers and other stable identifiers allow researchers to use sequence information with confidence over time.
- Sequence data is just the tip of the database iceberg

Next week: sequence comparisons

- **Sequence Alignment** is the fundamental algorithm of bioinformatics. Tomorrow we will begin to look at how to approach this problem
- We will see how comparing sequences allows us to understand and infer evolutionary relationships between sequences