BIOC8142

Unix II – Scripting, web clients, databases and formats

BIOC 8142
February 13, 2013
Bill Pearson wrp@virginia.edu 4-2818 Jordan 6-057

Goals of today's lecture:

- Creating simple bash scripts
- Survey of Bioinformatics databases (Ouellette)
  - Primary vs reference
  - Annotations and cross-references
  - Survey of file formats
- Scripts as web browsers

1

What should you do to reinforce the lecture material?

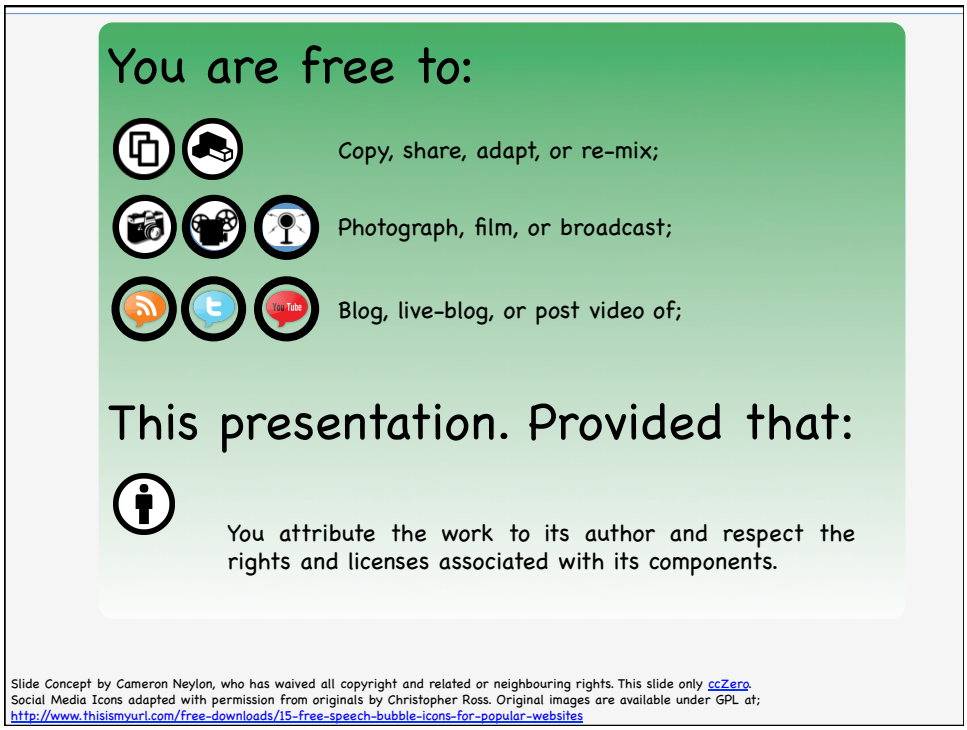
- Scripting the bash shell (Google "bash introduction", focus on variables, flow control)
  - tldp.org/LDP/abs/html/ (concise intro)
  - Learning the Bash Shell, 3rd edition (Ch 4 and 5) proquest.safaribooksonline.com/book/operating-systems-and-server-administration/unix/0596009658
- Bioinformatics databases:
  Pevsner (2004) "Bioinformatics and Functional Genomics 2nd ed" Wiley-Blackwell, Ch. 1 (on reserve, HSL)
- Web clients – `curl`, `wget` (`man curl`, `man wget`)

How will you be graded on this material:

- Homework on collab

2


Creative Commons
This page is available in the following languages:
Afrikaans Български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar italiano 日本語 한국어 Macedonian Melayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska
中文 華語 (台灣) isiZulu
creative commons
Attribution-Share Alike 2.5 Canada
You are free:
to Share — to copy, distribute and transmit the work
to Remix — to adapt the work
Free Cultural Works APPROVED FOR
Under the following conditions:
Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.
For any reuse or distribution, you must make clear to others the licence terms of this work.
Any of the above conditions can be waived if you get permission from the copyright holder.
The author's moral rights are retained in this licence.
Disclaimer
Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French
Learn how to distribute your work using this licence


You are free to:
Copy, share, adapt, or re-mix;
Photograph, film, or broadcast;
Blog, live-blog, or post video of;
This presentation. Provided that:
You attribute the work to its author and respect the rights and licenses associated with its components.
Slide Concept by Cameron Neylon, who has waived all copyright and related or neighbouring rights. This slide only ccZero.
Social Media Icons adapted with permission from originals by Christopher Ross. Original images are available under GPL at;
http://www.thisismyurl.com/free-downloads/15-free-speech-bubble-icons-for-popular-websites

Unix II – scripting, web clients, databases

- Scripting – putting commands in a file
  - bash commands:
    `for $file in ( *.fasta ); do ... Done`
  - Essential for reproducibility – your electronic lab notebook
  - Automation of repetitive tasks (run blast search using 20 files)
- Web clients – `curl/wget` – allow scripting of web access
  - Download a list of protein sequences using accessions
  - Homework – (a) do a blast search with tabular output; (b) extract accessions of hits; (c) download those sequences; (d) search with them

(bash) shell scripts

- files ending with `.sh` suffix
- shebang: `#!/bin/bash` or `#!/bin/sh`
- useful to capture (potentially long) history of UNIX commands into a reproducible analysis
  - you will always need to repeat your analysis
  - you will never remember all the necessary steps
- with some modification, your script can be made generic, and reusable for other data

shell scripts contain commands

```
franklin: 1 $ echo $PATH    # a simple command
/home/wrp/bin:/usr/local/bin:/bin:/usr/bin:.:/seqprg/bin

franklin: 2 $ echo_path.sh
                  # echo_path.sh contains "echo $PATH"
bash: ./echo_path.sh: Permission denied
                  # cannot execute because –rw-r–r--

franklin: 3 $ sh echo_path.sh # can execute with 'sh'
/home/wrp/bin:/usr/local/bin:/bin:/usr/bin:.:/seqprg/bin

franklin: 4 $ chmod +x echo_path.sh  # make executable

franklin: 5 $ echo_path.sh     # now it works
/home/wrp/bin:/usr/local/bin:/bin:/usr/bin:.:/seqprg/bin
```

(bash) shell variables

- Your unix session has two kinds of variables, `env` (environment) variables, and `SHELL` variables, refer to them with $NAME
  - Individual variables can be seen with '`echo`'
    `echo $PATH`
  - All environment variables are listed with `'env'`
- You can make your own variables for a command as well:
  `FILES=$(ls *.aa)`
  `echo $FILES`
  - shell variables never have a '`$`' on the left of the '=', and ALWAYS have a '`$`' on the right side.
  `NEW_FILES=$FILES`
- $SHELL variables are transient; to make them permanent, use:
  `export PATH=$PATH:/seqprg/bin`

(bash) shell flow control

- `for name in [...] ; do [...] ; done`
  - do something for each item in a list
- `if [...] ; then [...] ;`
  `elif [...]; then [...];`
  `else [...] fi`
  - specify behavior depending on conditions
- ';' are only necessary when putting multiple commands on one line.
  `for ... ; do ...; done`

Producing new filenames

```
franklin$ for f in *.aa;   # file glob (*)
> do
> n=$(basename $f .aa)  # $(command) makes output
                        into a string
> new=$n.new  # ${n} if no '.' or '/'
> new2="this${n}that"
> echo $f $new $new2
> done
gstm1_human.aa gstm1_human.new thisgstm1_humanthat
sequence.aa sequence.new thissequencethat
```

COMPUTATIONAL & COMPARATIVE GENOMICS: Understanding and Using Biological Databases

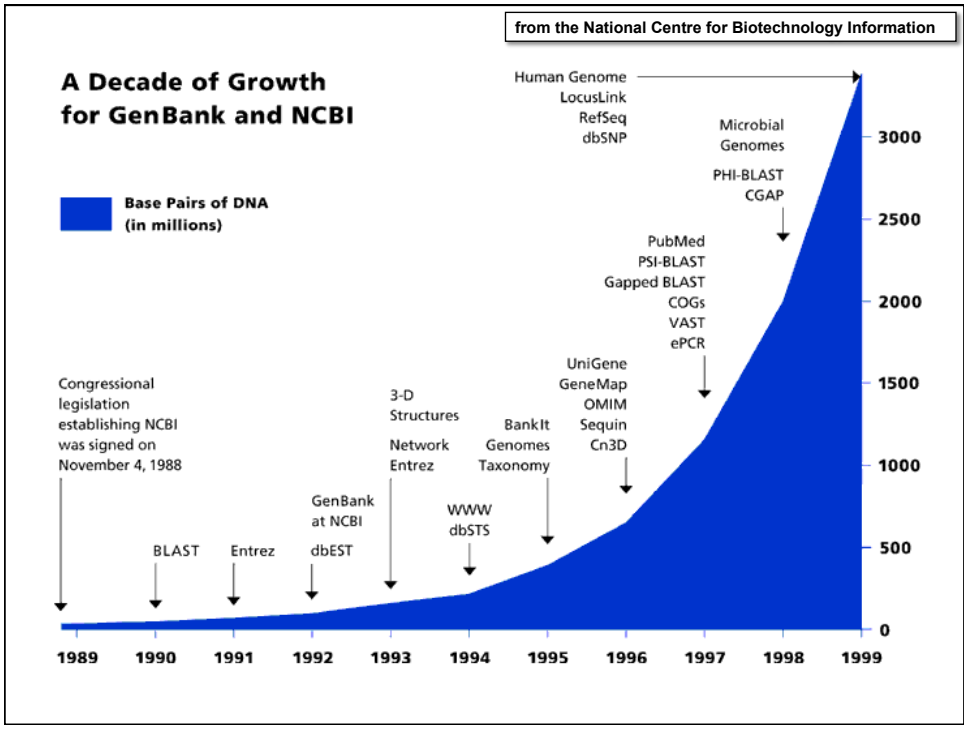
November 30th, 2012

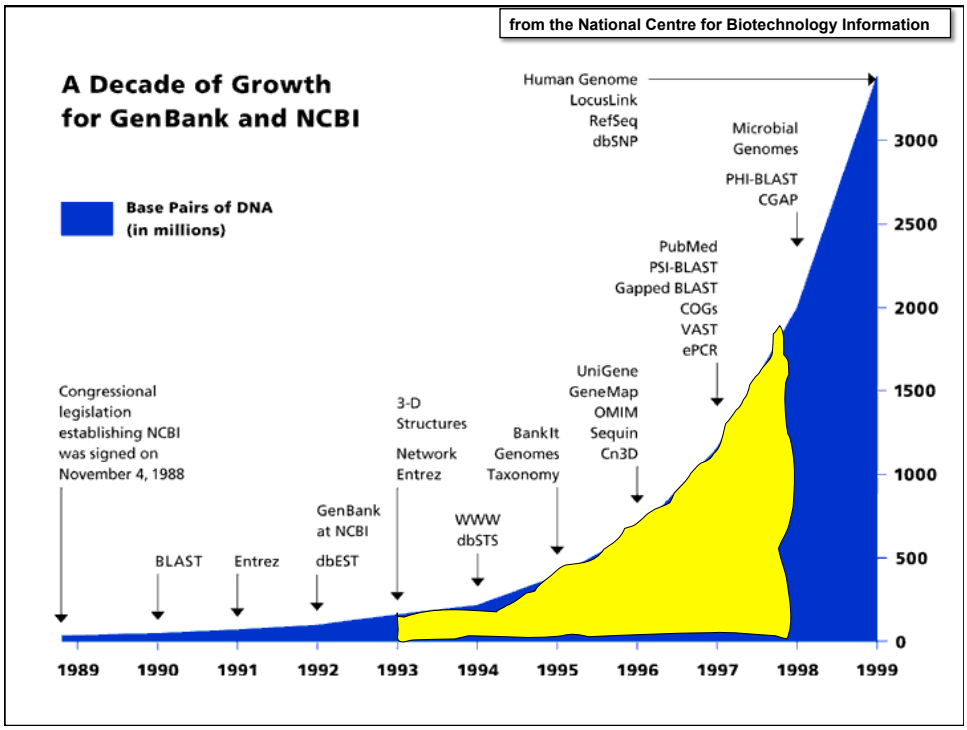
B.F. Francis Ouellette francis@oicr.on.ca

- Associate Director, Informatics and Biocomputing, Ontario Institute for Cancer Research, Toronto, ON
- Associate Professor, Department of Cell and Systems Biology, University of Toronto, Toronto, ON.

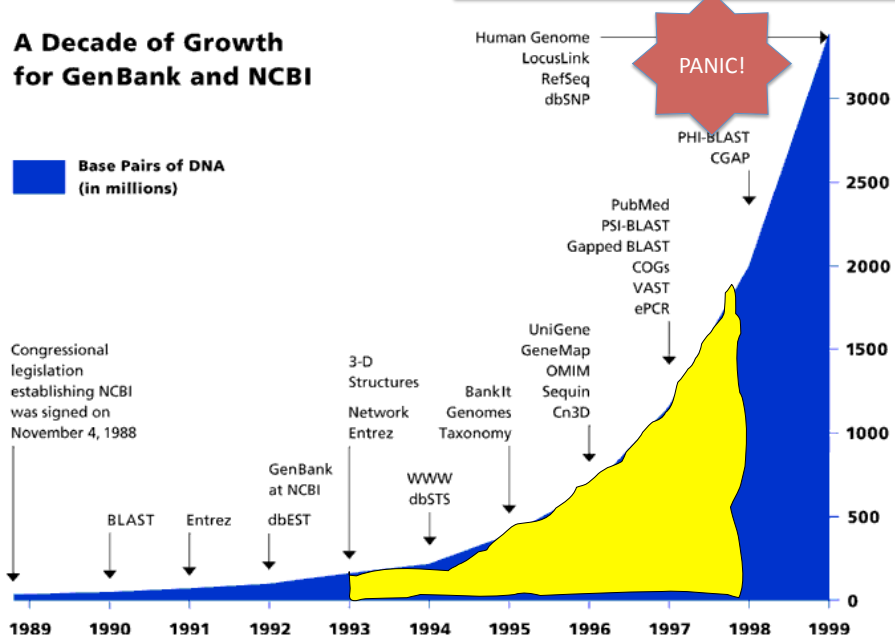
NCBI

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

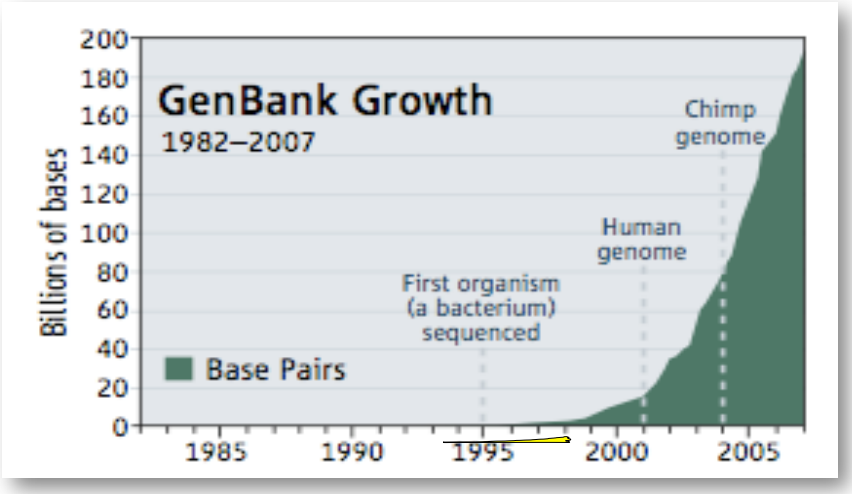

from the National Centre for Biotechnology Information
A Decade of Growth for GenBank and NCBI
Base Pairs of DNA (in millions)
Human Genome
LocusLink
RefSeq
dbSNP
Microbial Genomes
PHI-BLAST
CGAP
PubMed
PSI-BLAST
Gapped BLAST
COGs
VAST
ePCR
UniGene
GeneMap
OMIM
Sequin
Cn3D
BankIt
Genomes
Taxonomy
WWW
dbSTS
GenBank
at NCBI
dbEST
3-D
Structures
Network
Entrez
BLAST
Entrez
Congressional legislation establishing NCBI was signed on November 4, 1988
0
500
1000
1500
2000
2500
3000
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999


from the National Centre for Biotechnology Information
A Decade of Growth for GenBank and NCBI
Base Pairs of DNA (in millions)
Human Genome
LocusLink
RefSeq
dbSNP
Microbial Genomes
PHI-BLAST
CGAP
PubMed
PSI-BLAST
Gapped BLAST
COGs
VAST
ePCR
UniGene
GeneMap
OMIM
Sequin
Cn3D
BankIt
Genomes
Taxonomy
WWW
dbSTS
GenBank
at NCBI
dbEST
3-D
Structures
Network
Entrez
BLAST
Entrez
Congressional legislation establishing NCBI was signed on November 4, 1988
0
500
1000
1500
2000
2500
3000
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999

from the National Centre for Biotechnology Information



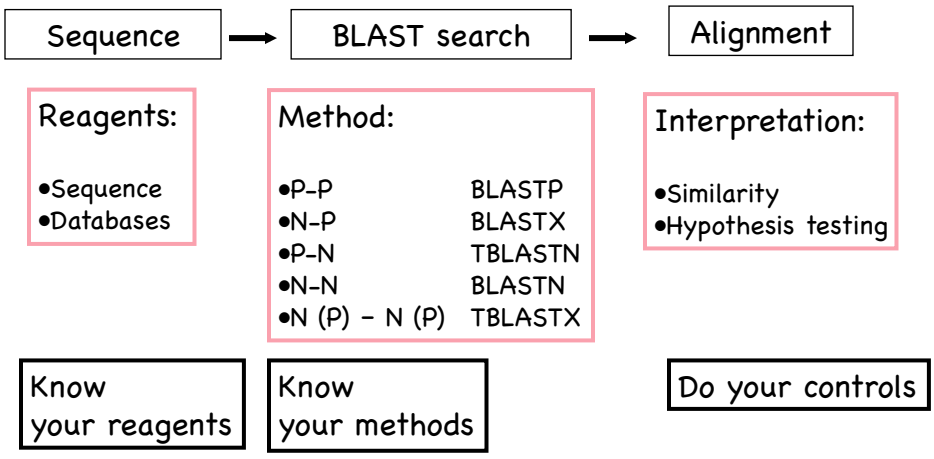
www.sciencemag.org SCIENCE VOL 319 21 MARCH 2008



Bioinformatics reagent: **Databases**

- Organized array of information
- Place where you put things in, and (if all is well) you should be able to get them out again.
- Resource for other databases and tools.
- Simplify the information space by specialization.
- Bonus: Allows you to make discoveries.
- Important question to ask:
  what is the data model?

Bioinformatics experiments:



18

Bioinformatics Citizenship: What it means, and what does it cost?

Users must help to keep public databases correct

Sir— With the continued growth of the public DNA sequence databases, and the recent addition of the 11,000,000,000th nucleotide to GenBank (including DDBJ, EMBL and GenBank), it is timely to assess how we use these databases.

GenBank is the archive of all publicly available DNA, RNA and protein sequences. Upon publication a new sequence and its annotations appear in it. Investigators use GenBank in many ways, most commonly for similarity searches such as BLAST; to retrieve records; and for sequence analysis, multiple sequence alignment or pattern finding search. Errors sometimes occur in GenBank, ranging from the trivial (incorrect postal codes), to the misleading (30 nucleotides of vector left on the ends of a record), to the mission-critical (a full length mRNA without a coding sequence (CDS) annotated on it). Also very common are incomplete references that prevent researchers from linking the GenBank record to the publication that refers to it first.

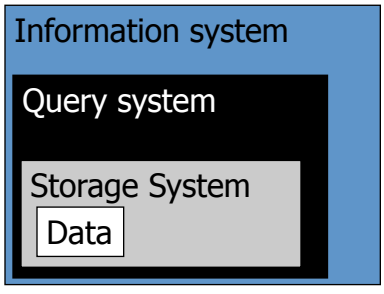
Over the years some people have chosen to report these errors, but in most cases they are left unmodified. An uncorrected 'discovered' error is one of the worst possible failings in GenBank, so if you discover an error, report it to the database (update@ncbi.nlm.nih.gov) and it should be rectified — although a follow-up is advised to make sure this gets done.

If you are a submitter, look at the record you submitted a few years ago: is it still correct? Was the citation ever updated? Take pride in the sequences that carry your name! Our ability to interpret genomes depends on all of these records being as accurate as possible. This is a task for all users of the databases.

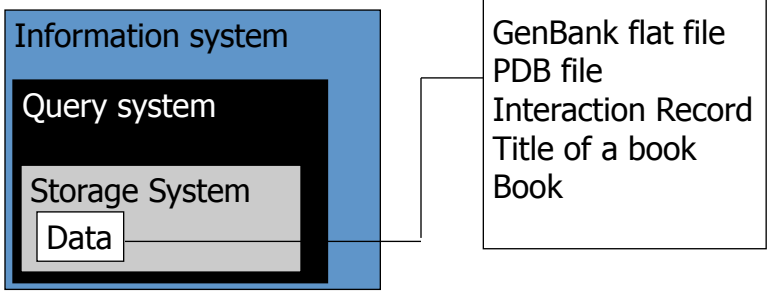
Francis Ouellette
Bioinformatics Core Facility, Centre for Molecular Medicine and Therapeutics, University of British Columbia, 950 West 28th Avenue, Vancouver, British Columbia V5Z 4H4, Canada

Nature **409**:452

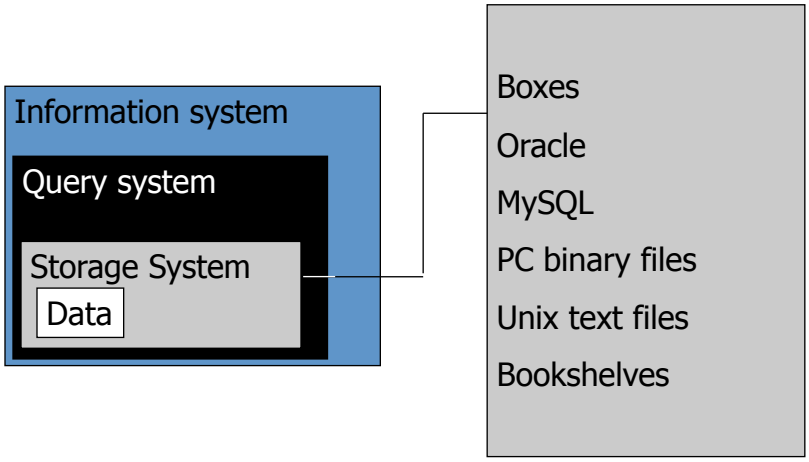
Databases



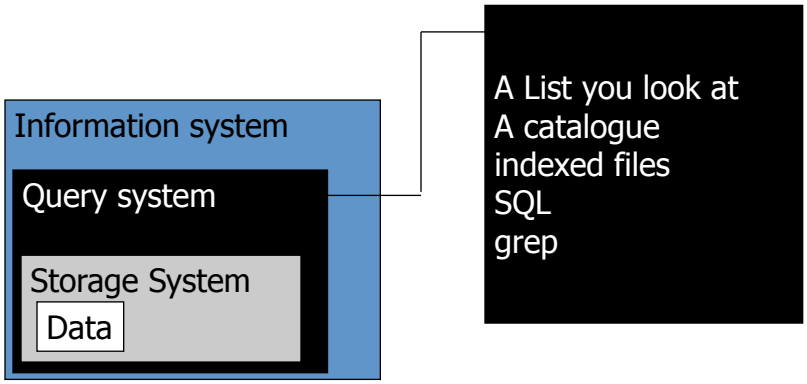
Databases



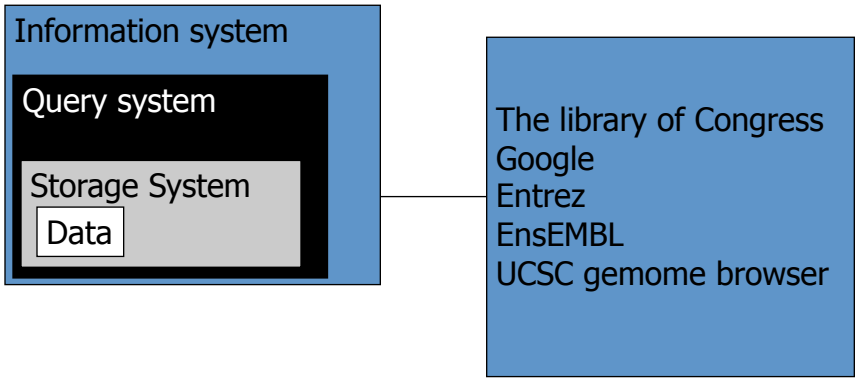
Databases



Databases



Databases



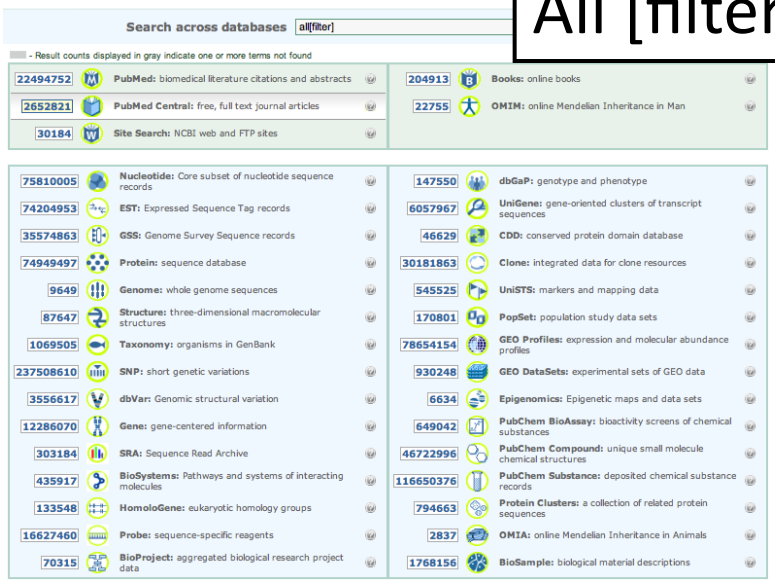
	Oct 15th, 2011
Nucleotide records	144,458,648
WGS records	68,330,215
Protein sequences	44,764,043
3D structures in PDB	76,973
Interactions and pathways	299,756
Human Unigene Cluster	122,727
Completed Genome projects	14,054
Different taxonomy Nodes	835,872
dbSNP records	149,211,539
Genes	9,096,508
RefSeq Genomic records	2,419,419
RefSeq RNA Records	2,679,762
RefSeq Protein Records	13,137,813
GenSAT images	106,414
GEO profiles	63,811,486
Homologene gene	128,030
PubChem compounds	28,485,889
PubMed records	21,298,405
Free PubMed records	3,466,553
PubMed Central records	2,303,372
OMIM records	21,992

	Oct 15th, 2011	Jul 17, 1999
Nucleotide records	144,458,648	4,456,822
WGS records	68,330,215	
Protein sequences	44,764,043	706,862
3D structures in PDB	76,973	9,780
Interactions and pathways	299,756	
Human Unigene Cluster	122,727	75,832
Completed Genome projects	14,054	10,870
Different taxonomy Nodes	835,872	52,889
dbSNP records	149,211,539	6,377
Genes	9,096,508	515
RefSeq Genomic records	2,419,419	
RefSeq RNA Records	2,679,762	
RefSeq Protein Records	13,137,813	
GenSAT images	106,414	
GEO profiles	63,811,486	
Homologene gene	128,030	
PubChem compounds	28,485,889	
PubMed records	21,298,405	10,372,886
Free PubMed records	3,466,553	
PubMed Central records	2,303,372	
OMIM records	21,992	10,695

	Oct 15th, 2011	Jul 17, 1999	fold differences
Nucleotide records	144,458,648	4,456,822	32
WGS records	68,330,215		
Protein sequences	44,764,043	706,862	63
3D structures in PDB	76,973	9,780	8
Interactions and pathways	299,756		
Human Unigene Cluster	122,727	75,832	2
Completed Genome projects	14,054	10,870	1
Different taxonomy Nodes	835,872	52,889	16
dbSNP records	149,211,539	6,377	23,398
Genes	9,096,508	515	17,663
RefSeq Genomic records	2,419,419		
RefSeq RNA Records	2,679,762		
RefSeq Protein Records	13,137,813		
GenSAT images	106,414		
GEO profiles	63,811,486		
Homologene gene	128,030		
PubChem compounds	28,485,889		
PubMed records	21,298,405	10,372,886	2
Free PubMed records	3,466,553		
PubMed Central records	2,303,372		
OMIM records	21,992	10,695	2

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

All [filter]



Feb, 2013

Formats

- DNA sequence (GenBank Flat Files)
- Protein Sequences
- Other formats to know about
  - FASTA
  - GFF3
  - XML

GenBank Flat File (GBFF)

```
LOCUS       JN675711               1704 bp    mRNA    linear   PLN 01-NOV-2011
DEFINITION  Prunus salicina mitogen-activated protein kinase 1 mRNA, complete
            cds.
ACCESSION   JN675711
VERSION     JN675711.1  GI:350528642
KEYWORDS    .
SOURCE      Prunus salicina
  ORGANISM  Prunus salicina
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicotyledons;
            rosids; fabids; Rosales; Rosaceae; Haloideae; Amygdaleae; Prunus.
REFERENCE   1  (bases 1 to 1704)
  AUTHORS   Jiang,C., Pan,D. and Chen,G.
  TITLE     Construction and Analysis of a Normalized Full-Length cDNA Library
            of Brown Prunus salicina
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1704)
  AUTHORS   Jiang,C., Pan,D. and Chen,G.
  TITLE     Direct Submission
  JOURNAL   Submitted (18-AUG-2011) Horticulture, Fujian Agriculture and
            Forestry University, Fuzhou, Fu Jian 350002, China
FEATURES             Location/Qualifiers
     source          1..1704
                     /organism="Prunus salicina"
                     /mol_type="mRNA"
                     /isolate="A"
                     /db_xref="taxon:88123"
     CDS             129..1249
                     /note="MAPK 1"
                     /codon_start=1
                     /product="mitogen-activated protein kinase 1"
                     /protein_id="AEQ28763.1"
                     /db_xref="GI:350528643"
                     /translation="MESSSASAGDNNIKGVPANGGRTVQVNVVGNFFEVSRKYVPPIR
                     PVGRGAYGIVCAAVNAETREVAIKKIGNAFDNRIDAKRTLREIKLLRHMDHENVIAI
                     KDIIRPPQRENFNDVYIVYELMDTDLHQIIRSNQPLMDDHCRYFLYQLLRGLKYVHSA
                     NVLHRDLKPSNLLHMNANCDLKICDFGLARTTSETDFMTEYVVTRWYRAPELLLNCSEY
                     TAAIDIWSVGCILGEIMTKKPLFPGKDYVHQLMLITELLGSPDDSSLGFLRSDNARRY
                     VRQLPQYPRQSFSAGFPMMSPGAVDLLEKMLVFDPNRRITVDEALCHPYLAPLHDINE
                     EPVCPMKFNFDFEQPSFTEEMIKELIWKESVRFNPDPIN"
ORIGIN
        1 ccattacggc ctagttacgg gggatcatta gtgctcttca tagctcctcc tctacaagtc
       61 ttagcctttt agcaccgaaa tcctaactgg ttctttgacc attccggatt cggatccctg
      121 gttaactatg gaatccagct ctgcttcagc aggtgatcac aatatcaaag gggtacctgc
      181 ccacggtgga cgctatgttc agtaccaatgt gtaacgttaac tctttgagg tttctaggaa
      241 gtacgtccct cccataaggc ccgtagggag aggtgcttat ggtattgttt gtgctgctgt
      301 gaatgctgag actcgtgagg aggttgccat taagaagatt ggtaatgcat tgacaacacg
      361 aatcgatgcc aagaggaatt tacgaggaaat taaacttctt cggcacatgg atcatgaaaa
      421 tgttattgcc atcaaagaca tcatacggcc tccacagaag gagaacttca atgatgtcta
      481 catgtattat gaattaatgg acactgatct tcacacagatt atacggtcca accaacctt
      541 gaacgatgat cattgtcggt actttctgta tcagttgtta agagggccta aatacgtaca
      601 ttcggcaaat gttttgcatc gtgatttgaa gcccagtaat ttgctcatga atgcgaattg
      661 cgacctaag aatggagatt ttggcctgc aaggacaaca tctgaaactg attcatatgc
      721 tgagtatgtt gttactcgtt ggtaccgagc accggagtta ctcctcaatt gttcggagta
      781 cactgcagca atcgatatat ggtctgtagg ttgcgattctt gggggaaatta tgaccagacg
      841 gcccctgttc cctggcaaag actatgtaca tcagctgaga ctcataacag agctcctagg
      901 ttcacctgat gactccagcc ttggatttt acgaagtgat aatgctcgaa gatatgtccg
      961 acaactacct cagtaccaa agcagagctt ctctgctgga ttcctcaata tgtctcctgg
     1021 cgctgttgat ttgctagaga agatgcttgt ctttgaccct aacaggcgca ttacagttga
     1081 tgaggctctt tgccaccct actggcacc tcttcatgat atcaacgagg agcctgtctg
     1141 cccgtatgcc ttcaatttg atttttgagca accatcctt acagaagaga acatcaagga
     1201 gttcatctgg agagaatctg taaagttcaa tccagatcca attcactgag gatcatgtta
     1261 taatctgtag tatgacttta gaaagtaatc ctaaaactga ttacaatgaa actctactgg
     1321 aagaaagaca ttataaaaac tgtgctggag agatggatt tcctgagaaa ttgtcctgag
     1381 atgtcattca ttcaacattca tcacgtcttt ggatgtctgt aattgtcaaa tgtgtaaaag
     1441 gagatcaaaa gctgaatctc aagtctgccg cgtcagttag aaccccaatc cctcatttgt
     1501 gtaggttgtt gcttttggtg tagacgtagg agctgtggtt agcggcctaa agtactatca
     1561 tgaccattaa cccagcactg ttgtagggcc cttgcggtct tgtgatcatc tctgttttaaa
     1621 aaggtaggtc tgaatataaa cggttgcctt ttttccaaaa gaaaaaaaaa aaaaaaaaaa
     1681 aaaaaaaaaa aaaaaaaaaa aaaa
//
```

Header
- Title
- Taxonomy
- Citation

Features (AA seq)

DNA Sequence

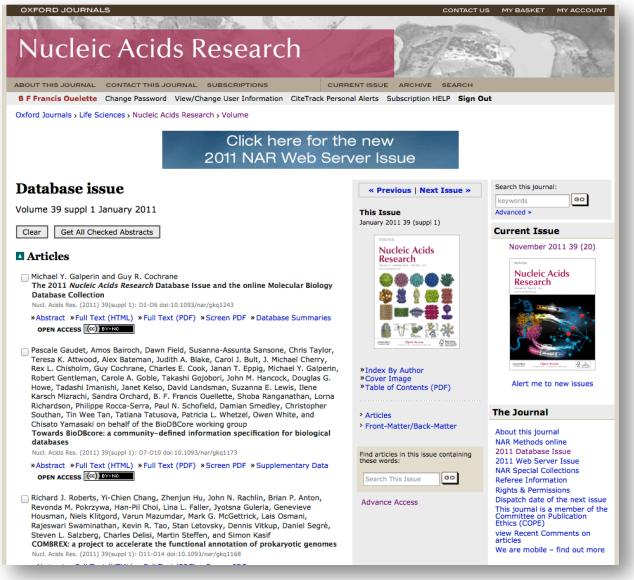
FASTA

```
>
MSEYQPSLFALNPMGFSPLDGSKSTNENVSASTSTAKPMVGQLIFDKFIKTEEDPI
IKQDTPSNLDFDFALPQTATAPDAKTVLPIPELDDAVVESFFSSSTDSTPMFEYEN
LEDNSKEWTSLFDNDIPVTTDDVSLADKAIESTEEVSLVPSNLEVSTTSFLPTPVL
EDAKLTQTRKVKKPNSVVKKSHHVGKDDESRLDHLGVVAYNRKQRSIPLSPIVPES
SDPAALKRARNTEAARRSRARKLQRMKQLEDKVEELLSKNYHLENEVARLKKLVGE
R
```

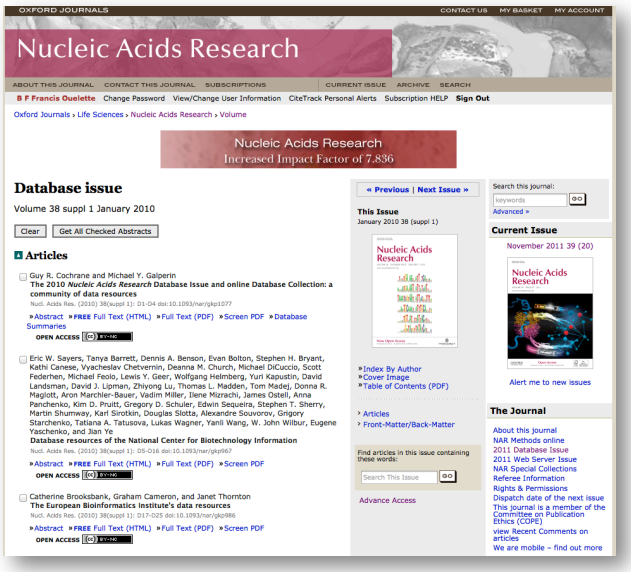
Databases

- Primary (archival)
  - GenBank/EMBL/DDBJ
  - UniProt
  - PDB
  - Medline (PubMed)
  - Intact
- Secondary (curated)
  - RefSeq
  - Taxon
  - UniProt
  - OMIM
  - SGD
  - Biosamples/ Bioprojects

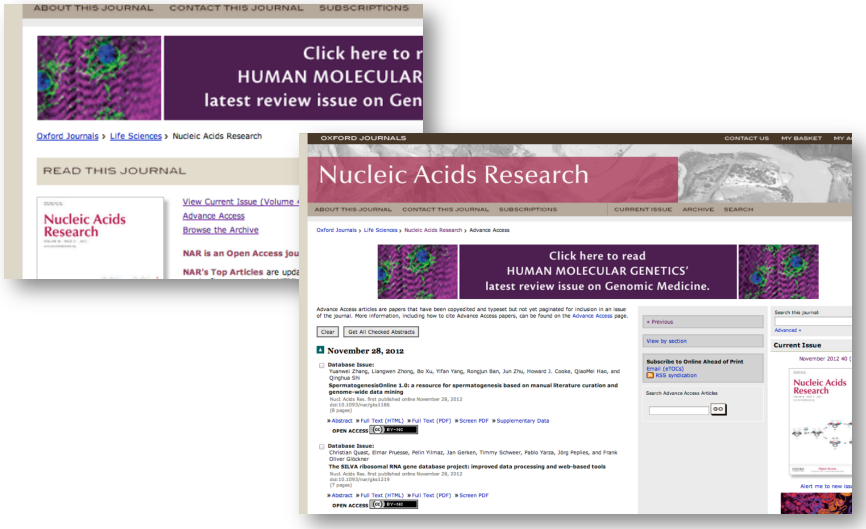
http://nar.oxfordjournals.org/content/39/suppl_1 January 2011



http://nar.oxfordjournals.org/content/38/suppl_1 January 2010



http://nar.oxfordjournals.org/



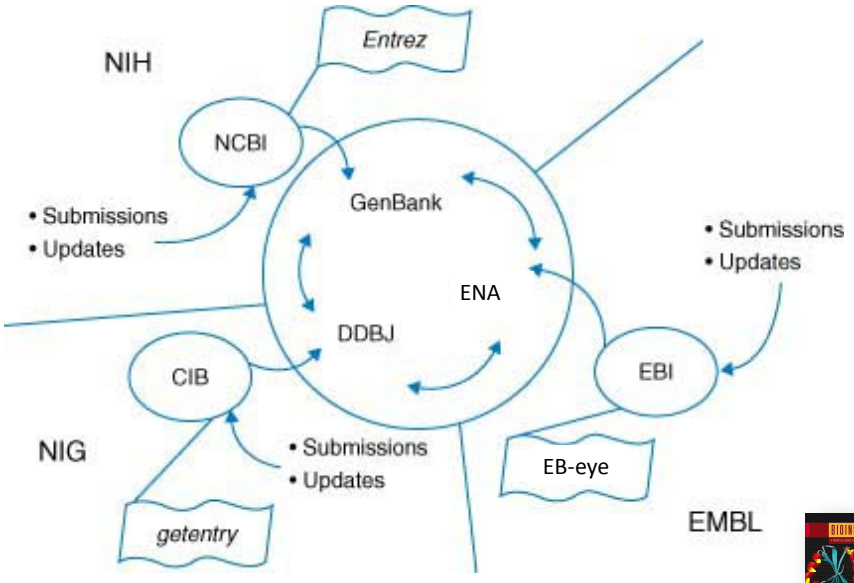
Sequence Databases

- Primary DNA (archive)
  - DDBJ/ENA/GenBank
- Primary protein (curated/automation)
  - UniProtKB
- Curated Databases (lots of human labour)
  - RefSeq (Genomic, mRNA and protein)
  - UniProtKB/SwissProt and neXtprot

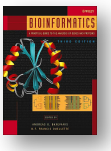
What is GenBank?

GenBank is the NIH genetic sequence database of all publicly available DNA and derived protein sequences, with annotations describing the biological information these records contain.

http://www.ncbi.nlm.nih.gov/genbank/
Benson et al., Nucleic Acids Res. 2012 (out in 2011)
http://www.ncbi.nlm.nih.gov/pubmed/22144687



Adapted from Rolf Apweiler



Types of files in GenBank

- From one-gene investigators
  - Often a very well annotated cDNA
  - A genomic segment from an new invertebrate
  - A mitochondria or virus
- From population/phylogenetic analysis
  - rRNA amplicon from environmental sampling
- From Genome Centers:
  - Gene expression:
    - Expressed Sequence Tags
    - Full Length Insert cDNA
    - TSA
  - Genome sequencing projects
    - HTG
    - CON

Organismal Divisions

		Used in which database?
BCT	Bacterial	DDBJ - GenBank
FUN	Fungal	ENA
HUM	Homo sapiens	DDBJ - ENA
INV	Invertebrate	all
MAM	Other mammalian	all
ORG	Organelle	ENA
PHG	Phage	all
PLN	Plant	all (not same data in all)
PRI	Primate (also see HUM)	all (not same data in all)
PRO	Prokaryotic	ENA
ROD	Rodent	all
SYN	Synthetic and chimeric	all
VRL	Viral	all
VRT	Other vertebrate	all

Functional Divisions

PAT Patent
EST Expressed Sequence Tags
TSA Transcriptome Shotgun Assembly
STS Sequence Tagged Site
GSS Genome Survey Sequence
HTG High Throughput Genome (unfinished)
HTC High throughput cDNA (unfinished)
CON Contig assembly instructions
ENV Environmental sampling methods

Organismal divisions:
BCT FUN INV MAM PHG PLN
PRI ROD SYN VRL VRT

Guiding Principals

In GenBank, records are grouped for various reasons: understand this is key to using and fully taking advantage of this database.

Identifiers

- You need identifiers which are stable through time
- Need identifiers which will always refer to specific sequences
- Need these identifiers to track history of **sequence** updates
- Also need feature and annotation identifiers (need to tract important things)
  - Genes
  - Transcripts
  - Proteins
  - (((Phenotype)))

LOCUS, Accession, NID and protein_id

LOCUS: Unique string of 10 letters and numbers in the database. Not maintained amongst databases, and is therefore a poor sequence identifier.

ACCESSION: A unique identifier to that record, citable entity; does not change when record is updated. A good record identifier, ideal for citation in publication.

VERSION: ID system where the accession and version play the same function as the accession and gi number.

Nucleotide gi: Geninfo identifier (gi), a unique integer which will change every time the sequence changes.

Protein gi: Geninfo identifier (gi), a unique integer which will change every time the sequence changes.

protein_id: Identifier which has the same structure and function as the nucleotide Accession.version numbers, but slightly different format.

LOCUS, Accession, gi and PID

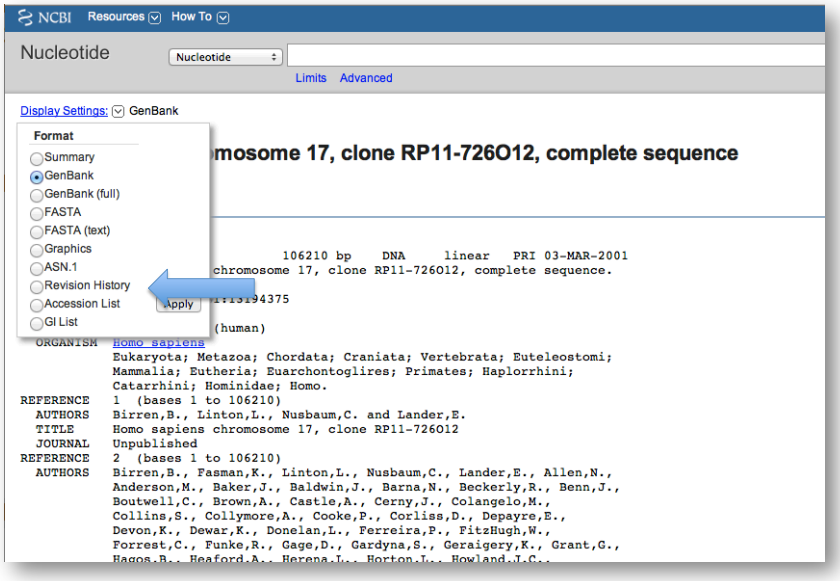
```
LOCUS       HSU40282     1789 bp    mRNA            PRI       21-MAY-1998
DEFINITION  Homo sapiens integrin-linked kinase (ILK) mRNA, complete cds.
ACCESSION   U40282
VERSION     U40282.1  GI:3150001
```

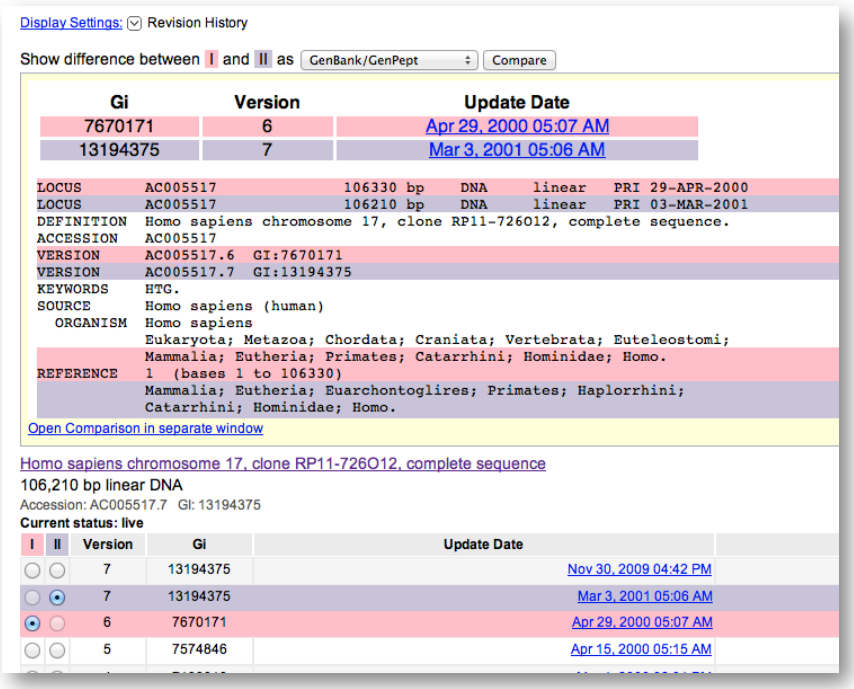
```
     LOCUS: HSU40282
 ACCESSION: U40282
   VERSION: U40282.1
        GI: 3150001
Protein gi: 3150002
protein_id: AAC16892.1
```



```
     CDS             157..1515
                     /gene="ILK"
                     /note="protein serine/threonine kinase"
                     /codon_start=1
                     /product="integrin-linked kinase"
                     /protein_id="AAC16892.1"
                     /db_xref="GI:3150002"
```

Lecture 1.3





Display Settings: Revision History

Show difference between I and II as GenBank/GenPept Compare

	Gi	Version	Update Date
I	7670171	6	Apr 29, 2000 05:07 AM
II	13194375	7	Mar 3, 2001 05:06 AM

```
LOCUS       AC005517               106330 bp    DNA     linear   PRI 29-APR-2000
LOCUS       AC005517               106210 bp    DNA     linear   PRI 03-MAR-2001
DEFINITION  Homo sapiens chromosome 17, clone RP11-726O12, complete sequence.
ACCESSION   AC005517
VERSION     AC005517.6  GI:7670171
VERSION     AC005517.7  GI:13194375
KEYWORDS    HTG.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 106330)
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
```

Open Comparison in separate window

Homo sapiens chromosome 17, clone RP11-726O12, complete sequence

106,210 bp linear DNA

Accession: AC005517.7 GI: 13194375

Current status: live

I	II	Version	Gi	Update Date
○	○	7	13194375	Nov 30, 2009 04:42 PM
○	●	7	13194375	Mar 3, 2001 05:06 AM
●	○	6	7670171	Apr 29, 2000 05:07 AM
○	○	5	7574846	Apr 15, 2000 05:15 AM

Accession number "space"

- GenBank:
  - 1+5 (L12345, U00001)
  - 2+6 (AF000001, AC000003)
- WGS (*Not distributed with* GenBank)
  - 4+2+6 (AAAA01000001, AAAD01000001)
- Protein:
  - 1+5 or 3+5
- All have "accession.version"

Secondary Accession Numbers

- When you 'retire' accession numbers, these often are put in the secondary accession number space. (e.g GenBank Accession number L05146)
- With the removal of sequence length limits, GenBank will now allow continuous ranges of secondary accessions.
- As of GenBank Release 146.0 (February 2005), it is legal to represent continuous ranges of secondary accessions by a start accession, a dash character, and an end accession. (e.g. for the *E. coli* genome)

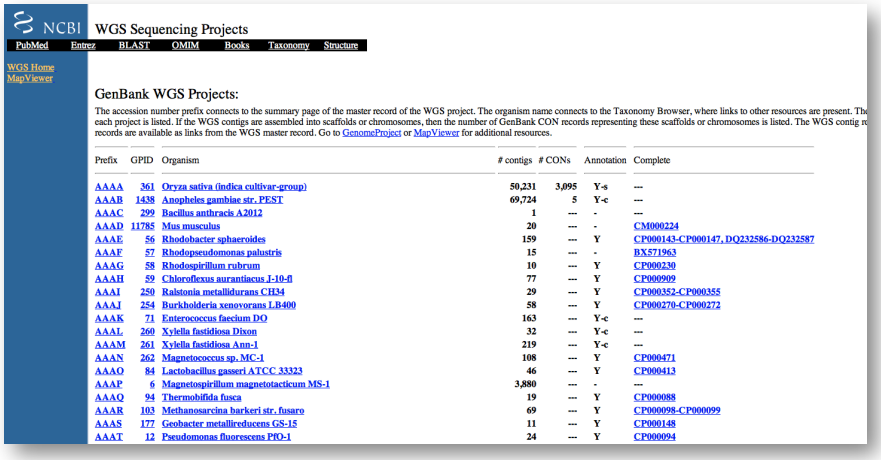
ACCESSION U00096 AE000111-AE000510

```
LOCUS       U00096               4639675 bp    DNA     circular BCT 01-SEP-2011
DEFINITION  Escherichia coli str. K-12 substr. MG1655, complete genome.
ACCESSION   U00096 AE000111-AE000510
VERSION     U00096.2  GI:48994873
DBLINK      Project: 225
KEYWORDS    .
SOURCE      Escherichia coli str. K-12 substr. MG1655
  ORGANISM  Escherichia coli str. K-12 substr. MG1655
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (bases 1 to 4639675)
  AUTHORS   Blattner,F.R., Plunkett,G. III, Bloch,C.A., Perna,N.T., Burland,V.,
            Riley,M., Collado-Vides,J., Glasner,J.D., Rode,C.K., Mayhew,G.F.,
            Gregor,J., Davis,N.W., Kirkpatrick,H.A., Goeden,M.A., Rose,D.J.,
            Mau,B. and Shao,Y.
  TITLE     The complete genome sequence of Escherichia coli K-12
  JOURNAL   Science 277 (5331), 1453-1474 (1997)
   PUBMED   9278503
```

WGS: Whole Genome Shotgun (Not in GenBank release, not shared with ENA/DDBJ)

- Contigs from ongoing Whole Genome Shotgun sequencing projects
- The nucleotides from WGS projects go into the BLAST 'wgs' database, whereas the proteins go into the BLAST nr database.
- More info, and how to submit to this division: http://www.ncbi.nlm.nih.gov/Genbank/wgs.html
- Accession format is 4+2+6

http://www.ncbi.nlm.nih.gov/projects/WGS/WGSprojectlist.cgi



NCBI WGS Sequencing Projects

PubMed Entrez BLAST OMIM Books Taxonomy Structure

WGS Home
MapViewer

GenBank WGS Projects:

The accession number prefix connects to the summary page of the master record of the WGS project. The organism name connects to the Taxonomy Browser, where links to other resources are present. The each project is listed. If the WGS contigs are assembled into scaffolds or chromosomes, then the number of GenBank CON records representing these scaffolds or chromosomes is listed. The WGS contig r records are available as links from the WGS master record. Go to GenomeProject or MapViewer for additional resources.

Prefix	GPID	Organism	# contigs	# CONs	Annotation	Complete
AAAA	361	Oryza sativa (indica cultivar-group)	50,231	3,095	Y-s	---
AAAB	1438	Anopheles gambiae str. PEST	69,724	5	Y-c	---
AAAC	299	Bacillus anthracis A2012	1	---	-	---
AAAD	11785	Mus musculus	20	---	-	CM000224
AAAE	56	Rhodobacter sphaeroides	159	---	Y	CP000143-CP000147, DQ232586-DQ232587
AAAF	57	Rhodopseudomonas palustris	15	---	-	BX571963
AAAG	58	Rhodospirillum rubrum	10	---	Y	CP000230
AAAH	59	Chloroflexus aurantiacus J-10-fl	77	---	Y	CP000909
AAAI	250	Ralstonia metallidurans CH34	29	---	Y	CP000352-CP000355
AAAJ	254	Burkholderia xenovorans LB400	58	---	Y	CP000270-CP000272
AAAK	71	Enterococcus faecium DO	163	---	Y-c	---
AAAL	260	Xylella fastidiosa Dixon	32	---	Y-c	---
AAAM	261	Xylella fastidiosa Ann-1	219	---	Y-c	---
AAAN	262	Magnetococcus sp. MC-1	108	---	Y	CP000471
AAAO	84	Lactobacillus gasseri ATCC 33323	46	---	Y	CP000413
AAAP	6	Magnetospirillum magnetotacticum MS-1	3,880	---	-	---
AAAQ	94	Thermobifida fusca	19	---	Y	CP000088
AAAR	103	Methanosarcina barkeri str. fusaro	69	---	Y	CP000098-CP000099
AAAS	177	Geobacter metallireducens GS-15	11	---	Y	CP000148
AAAT	12	Pseudomonas fluorescens PfO-1	24	---	Y	CP000094

52

WGS record (not in GenBank)

```
LOCUS       AACO02000001           55046 bp    DNA     linear   PLN 03-OCT-2006
DEFINITION  Cryptococcus neoformans var. grubii H99 contg1.1, whole genome
            shotgun sequence.
ACCESSION   AACO02000001 AACO02000000
VERSION     AACO02000001.1  GI:115505304
KEYWORDS    WGS.
SOURCE      Cryptococcus neoformans var. grubii H99
  ORGANISM  Cryptococcus neoformans var. grubii H99
            Eukaryota; Fungi; Basidiomycota; Hymenomycetes;
            Heterobasidiomycetes; Tremellomycetidae; Tremellales; Tremellaceae;
            Filobasidiella.
```

```
REFERENCE   3  (bases 1 to 55046)
  AUTHORS   Birren,B., Lander,E., Galagan,J., Nusbaum,C., Devon,K., Ma,L.-J.,
            Jaffe,D., Butler,J., Alvarez,P., Gnerre,S., Grabherr,M., Kleber,M.
            Mauceli,E., Brockman,W., Rounsley,S., Young,S., LaButti,K.,
            Pushparaj,V., DeCaprio,D., Crawford,M., Koehrsen,M., Engels,R.,
            Montgomery,P., Pearson,M., Howarth,C., Larson,L., Luoma,S.,
            White,J., Kodira,C., O'Leary,S., Yandava,C., Zeng,Q., Alvarado,L.,
            Dietrich,F., Heitman,J. and Kronstad,J.
  CONSRTM   The Broad Institute Genome Sequencing Platform
  TITLE     Direct Submission
  JOURNAL   Submitted (26-SEP-2006) Broad Institute of MIT and Harvard, 320
            Charles Street, Cambridge, MA 02141, USA
FEATURES             Location/Qualifiers
     source          1..55046
                     /organism="Cryptococcus neoformans var. grubii H99"
                     /mol_type="genomic DNA"
                     /strain="H99"
                     /variety="grubii"
                     /serotype="A"
                     /db_xref="taxon:235443"
ORIGIN
        1 gtcaagagat aaacagccat gctcaccata cacagactat gatccgtcag cattgtcttt
       61 tgtcacgagg ataaacatgg acgttaccgc aaactcctct ccatccacac cctatattct
      121 ccatcgccag cctttccatc ttgcccacct ctctcgccat cctgctttcg acaccgtctg
      181 ctctttctggc gtccttccgc actctgttat cctttccctt ccggctgttt gactatcgac
      241 aacagtcagc gagagtaaca ctcagagcaa cggcgacaaa cgaatcagca actcacagtc
      301 tgtacagcgt actatatctc cagcatccac cgccctgccc tccgccgtgt accatacctc
      361 ctccctgctg tctcatccac acgcaaccaa acgttgcttg cttgcttgcg tcaagagtag
```

Sequences NOT in GenBank

- WGS: whole genome shotgun
- TPA: third party annotations
- SNPs
- SAGE tags (serial analysis of gene expression)
- RefSeq (Genomic, mRNA, or protein)
- Consensus sequences

What is UniProtKB?

UniProt is a protein sequence database that is the result of a merge from SWISS-PROT and PIR and is funded by the NIH, EMBL and the Swiss Gov't. It is the main distributed, annotated, and curated protein sequence database. Data in UniProt is derived from coding sequence annotations in ENA (GenBank/ENA/DDBJ) nucleic acid sequence data, and from sequences in PIR and SP. UniProt is a Flat-File database just like GenBank or ENA/EMBL

- http://www.uniprot.org/
- http://database.oxfordjournals.org/content/2011/bar009.long


DATABASE

Database, Vol. 2011, Article ID bar009, doi:10.1093/database/bar009

Original article

UniProt Knowledgebase: a hub of integrated protein data

Michele Magrane[1,*] and UniProt Consortium[1,2,3]

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK, [2]Swiss Institute of Bioinformatics, Centre Medical Universitaire, 1 rue Michel Servet, 1211 Geneva 4, Switzerland, [3]Protein Information Resource, Georgetown University Medical Center, 3300 Whitehaven St. NW, Suite 1200, Washington, DC 20007, University of Delaware, 15 Innovation Way, Suite 205, Newark, DE 19711, USA

*Corresponding author: Tel: +44 (0)1223 494 696; Fax: +44 (0)1223 494 468; Email: magrane@ebi.ac.uk

Submitted 24 November 2010; Accepted 10 March 2011

Uniprot Curation

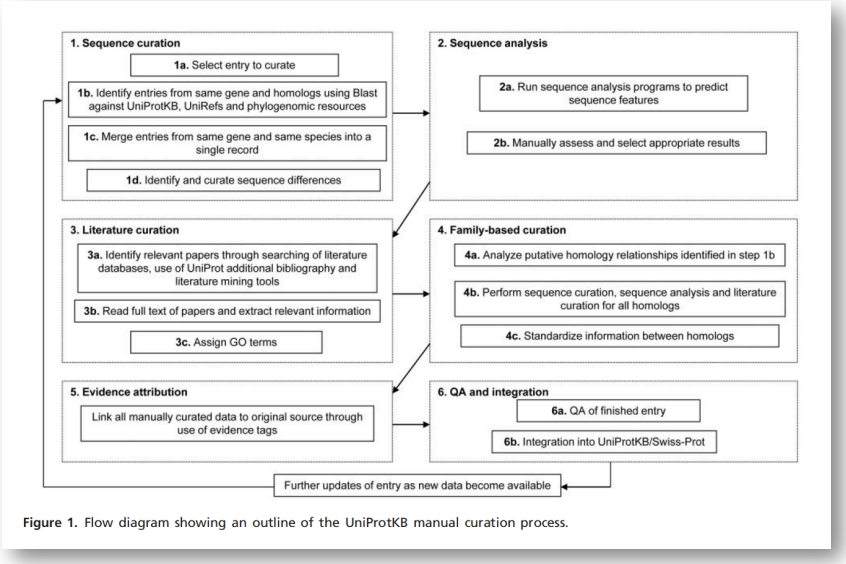
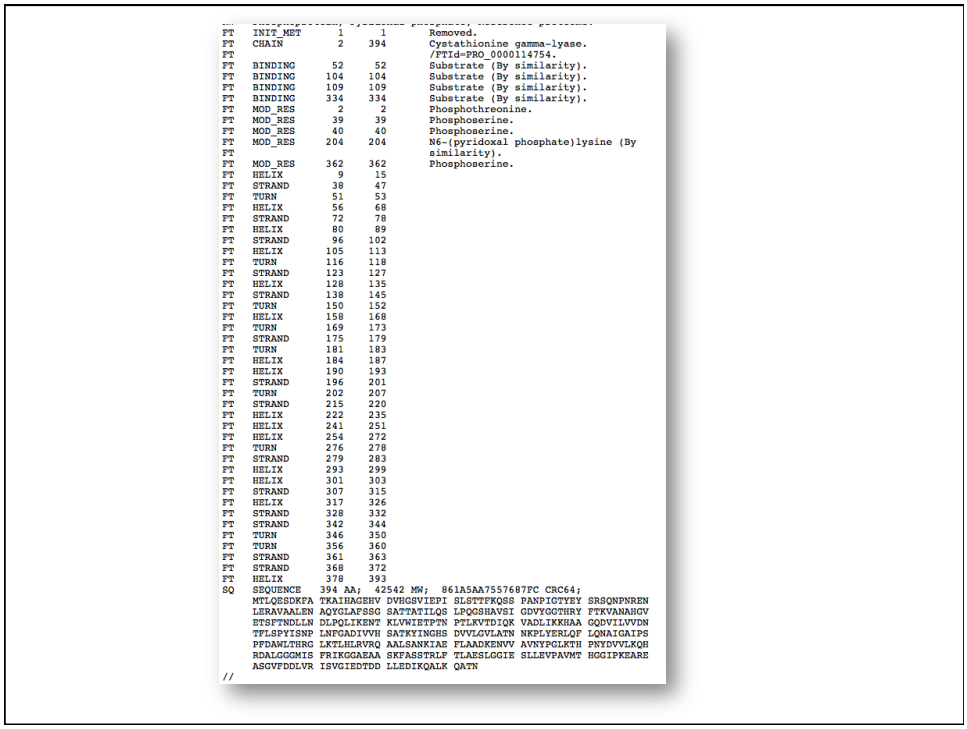


Figure 1. Flow diagram showing an outline of the UniProtKB manual curation process.

```
ID   CYS3_YEAST              Reviewed;         394 AA.
AC   P31373; D6VPK6;
DT   01-JUL-1993, integrated into UniProtKB/Swiss-Prot.
DT   23-JAN-2007, sequence version 2.
DT   19-OCT-2011, entry version 115.
DE   RecName: Full=Cystathionine gamma-lyase;
DE            EC=4.4.1.1;
DE   AltName: Full=Gamma-cystathionase;
DE   AltName: Full=Sulfur transfer protein 1;
GN   Name=CYS3; Synonyms=CYI1, STR1; OrderedLocusNames=YAL012W;
GN   ORFNames=FUN35;
OS   Saccharomyces cerevisiae (strain ATCC 204508 / S288c) (Baker's yeast).
OC   Eukaryota; Fungi; Dikarya; Ascomycota; Saccharomycotina;
OC   Saccharomycetes; Saccharomycetales; Saccharomycetaceae; Saccharomyces.
OX   NCBI_TaxID=559292;
RN   [1]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA], AND PARTIAL PROTEIN SEQUENCE.
RX   MEDLINE=92250430; PubMed=1577698;
RA   Ono B., Tanaka K., Naito K., Heike C., Shinoda S., Yamamoto S.,
RA   Ohmori S., Oshima T., Toh-e A.;
RT   "Cloning and characterization of the CYS3 (CYI1) gene of Saccharomyces
RT   cerevisiae.";
RL   J. Bacteriol. 174:3339-3347(1992).
RN   [2]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA], AND CHARACTERIZATION.
RC   STRAIN=DBY939;
RX   MEDLINE=93328685; PubMed=8335636;
RA   Yamagata S., D'Andrea R.J., Fujisaki S., Isaji M., Nakamura K.;
RT   "Cloning and bacterial expression of the CYS3 gene encoding
RT   cystathionine gamma-lyase of Saccharomyces cerevisiae and the
RT   physicochemical and enzymatic properties of the protein.";
RL   J. Bacteriol. 175:4800-4808(1993).
RN   [3]
RP   NUCLEOTIDE SEQUENCE [GENOMIC DNA].
RC   STRAIN=ATCC 204511 / S288c / AB972;
RX   MEDLINE=93289814; PubMed=8511966; DOI=10.1002/yea.320090406;
RA   Barton A.B., Kaback D.B., Clark M.W., Keng T., Ouellette B.F.F.,
RA   Storms R.K., Zeng B., Zhong W.W., Fortin N., Delaney S., Bussey H.;
RT   "Physical localization of yeast CYS3, a gene whose product resembles
RT   the rat gamma-cystathionase and Escherichia coli cystathionine gamma-
RT   synthase enzymes.";
RL   Yeast 9:363-369(1993).
```



```
FT   INIT_MET      1      1       Removed.
FT   CHAIN         2    394       Cystathionine gamma-lyase.
FT                                /FTId=PRO_0000114754.
FT   BINDING      52     52       Substrate (By similarity).
FT   BINDING     104    104       Substrate (By similarity).
FT   BINDING     109    109       Substrate (By similarity).
FT   BINDING     334    334       Substrate (By similarity).
FT   MOD_RES       2      2       Phosphothreonine.
FT   MOD_RES      39     39       Phosphoserine.
FT   MOD_RES      40     40       Phosphoserine.
FT   MOD_RES     204    204       N6-(pyridoxal phosphate)lysine (By
FT                                similarity).
FT   MOD_RES     362    362       Phosphoserine.
FT   HELIX         9     15
FT   STRAND       38     47
FT   TURN         51     53
FT   HELIX        56     68
FT   STRAND       72     78
FT   HELIX        80     89
FT   STRAND       96    102
FT   HELIX       105    113
FT   TURN        116    118
FT   STRAND      123    127
FT   HELIX       128    135
FT   STRAND      138    145
FT   TURN        150    152
FT   HELIX       158    168
FT   TURN        169    173
FT   STRAND      175    179
FT   TURN        181    183
FT   HELIX       184    187
FT   HELIX       190    193
FT   STRAND      196    201
FT   TURN        202    207
FT   STRAND      215    220
FT   HELIX       222    235
FT   HELIX       241    251
FT   HELIX       254    272
FT   TURN        276    278
FT   STRAND      279    283
FT   HELIX       293    299
FT   HELIX       301    303
FT   STRAND      307    315
FT   HELIX       317    326
FT   STRAND      328    332
FT   STRAND      342    344
FT   TURN        346    350
FT   TURN        356    360
FT   STRAND      361    363
FT   STRAND      368    372
FT   HELIX       378    393
SQ   SEQUENCE   394 AA;  42542 MW;  861A5AA7557687FC CRC64;
     MTLQESDKFA TKAIHAGEHV DVHGSVIEPI SLSTTFKQSS PANPIGTYEY SRSQNPNREN
     LERAVAALEN AQYGLAFSSG SATTATILQS LPQGSHAVSI GDVYGGTHRY FTKVANAHGV
     ETSFTNDLLN DLPQLIKENT KLVWIETPTN PTLKVTDIQK VADLIKKHAA QDVILVVDN
     TFLSPYISNP LNFGADIVVH SATKYINGHS DVVLGVLATN KPLYERLQFL QNAIGAIPS
     PFDAWLTHRG LKTLHLRVRQ AALSANKIAE FLAADKENVV AVNYPGLKTH PNYDVVLKQH
     RDALGGGMIS FRIKGGAEAA SKFASSTRLF TLAESLGGIE SLLEVPAVMT HGGIPKEARE
     ASGVFDDLVR ISVGIEDTDD LLEDIKQALK QATN
//
```

UniprotKB/SwissProt

```
ID   CYS3_YEAST     STANDARD;      PRT;   393 AA.
AC   P31373;
DT   01-JUL-1993 (REL. 26, CREATED)
DT   01-JUL-1993 (REL. 26, LAST SEQUENCE UPDATE)
DT   01-NOV-1995 (REL. 32, LAST ANNOTATION UPDATE)
DE   CYSTATHIONINE GAMMA-LYASE (EC 4.4.1.1) (GAMMA-CYSTATHIONASE).
GN   CYS3 OR CYI1 OR STR1 OR YAL012W OR FUN35.
OS   TAXONOMY
OC   SACCHAROMYCETACEAE; SACCHAROMYCES.

RX   CITATION
CC   -!- CATALYTIC ACTIVITY: L-CYSTATHIONINE + H(2)O = L-CYSTEINE +
CC       NH(3) + 2-OXOBUTANOATE.
CC   -!- COFACTOR: PYRIDOXAL PHOSPHATE.
CC   -!- PATHWAY: FINAL STEP IN THE TRANS-SULFURATION PATHWAY SYNTHESIZING
CC       L-CYSTEINE FROM L-METHIONINE.
CC   -!- SUBUNIT: HOMOTETRAMER.
CC   -!- SUBCELLULAR LOCATION: CYTOPLASMIC.
CC   -!- SIMILARITY: BELONGS TO THE TRANS-SULFURATION ENZYMES FAMILY.
CC   ---------------------------------------------------------------------------
CC   DISCLAMOR
CC   ---------------------------------------------------------------------------

DR   DATABASE cross-reference
KW   CYSTEINE BIOSYNTHESIS; LYASE; PYRIDOXAL PHOSPHATE.
FT   INIT_MET      0      0
FT   BINDING     203    203       PYRIDOXAL PHOSPHATE (BY SIMILARITY).
SQ   SEQUENCE   393 AA;  42411 MW;  55BA2771 CRC32;
     TLQESDKFAT KAIHAGEHVD VHGSVIEPIS LSTTFKQSSP ANPIGTYEYS RSQNPNRENL
     ERAVAALENA QYGLAFSSGS ATTATILQSL PQGSHAVSIG DVYGGTHRYF TKVANAHGVE
     TSFTNDLLND LPQLIKENTK LVWIETPTNP TLKVTDIQKV ADLIKKHAAG QDVILVVDNT
     FLSPYISNPL NFGADIVVHS ATKYINGHSD VVLGVLATNN KPLYERLQFL QNAIGAIPSP
     FDAWLTHRGL KTLHLRVRQA ALSANKIAEF LAADKENVVA VNYPGLKTHP NYDVVLKQHR
     DALGGGMISF RIKGGAEAAS KFASSTRLFT LAESLGGIES LLEVPAVMTH GGIPKEAREA
     SGVFDDLVRI SVGIEDTDDL LEDIKQALKQ ATN
//
```

UniProtKB

UniProt incorporates:

- Function of the protein
- Post-translational modification
- Domains and sites.
- Secondary structure.
- Quaternary structure.
- Similarities to other proteins;
- Diseases associated with deficiencies in the protein
- Sequence conflicts, variants, etc.

In closing ...

- Often only use FASTA files (eg for BLAST)
- Using any sequence where the coordinates are important, need an accession.version OR a gi number
- GBFF and neXtprot are simply human readable versions of these records
- GBFF have become a vehicle for a lot more information than they where meant to do when they were created!
- Keep in mind that GenBank is DNA centric and is a poor vehicle for protein and mRNA expression/interaction information: NCBI (and others) have other databases for these entities.
- All databases I mentioned today are fully “open” …

http://goo.gl/bdA1q

Users must help to keep public databases correct

Sir—With the continued growth of the public DNA sequence databases, and the recent addition of the 11,000,000,000th nucleotide to GenBank (including DDBJ, EMBL and GenBank), it is timely to assess how we use these databases.

GenBank is the archive of all publicly available DNA, RNA and protein sequences. Upon publication a new sequence and its annotations appear in it. Investigators use GenBank in many ways, most commonly for similarity searches such as BLAST; to retrieve records; and for sequence analysis, multiple sequence alignment or pattern finding search. Errors sometimes occur in GenBank, ranging from the trivial (incorrect postal codes), to the misleading (30 nucleotides of vector left on the ends of a record), to the mission-critical (a full length mRNA without a coding sequence (CDS) annotated on it). Also very common are incomplete references that prevent researchers from linking the GenBank record to the publication that refers to it first.

Over the years some people have chosen to report these errors, but in most cases they are left unmodified. An uncorrected 'discovered' error is one of the worst possible failings in GenBank, so if you discover an error, report it to the database (update@ncbi.nlm.nih.gov) and it should be rectified—although a follow-up is advised to make sure this gets done.

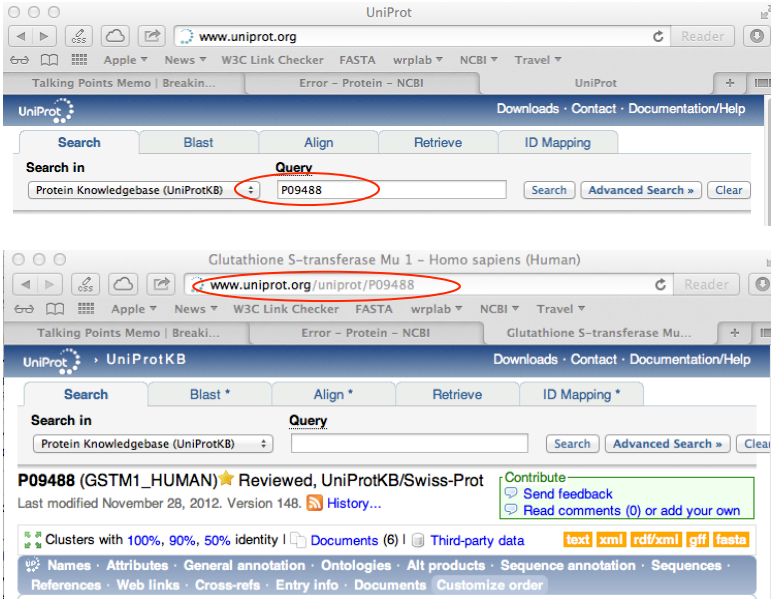
If you are a submitter, look at the record you submitted a few years ago: is it still correct? Was the citation ever updated? Take pride in the sequences that carry your name! Our ability to interpret genomes depends on all of these records being as accurate as possible. This is a task for all users of the databases.

Francis Ouellette
Bioinformatics Core Facility, Centre for Molecular Medicine and Therapeutics, University of British Columbia, 950 West 28th Avenue, Vancouver, British Columbia V5Z 4H4, Canada

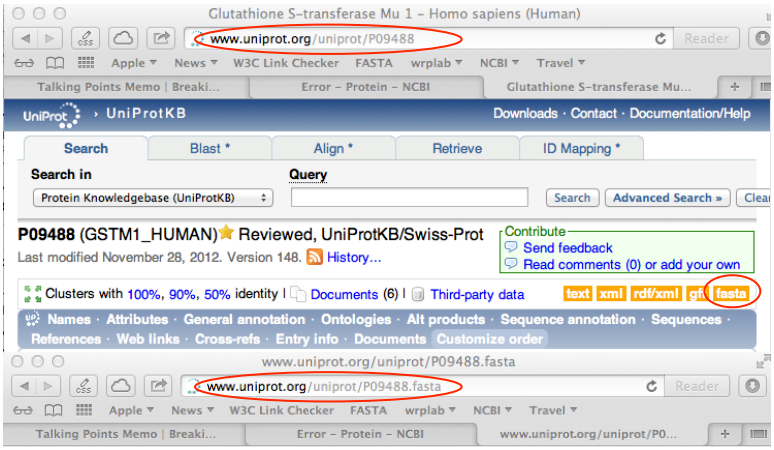
Scripting from the WWW: `wget/curl`

- Most bioinformatic analyses require resources from the web, e.g. sequences, domain information, datasets, etc.
  - The NCBI and EBI resources are usually scriptable; e.g. write a script that takes a set of accessions from a file and get the sequences
  - Often all that is required is to recognize the URL of the information desired
    http://www.ncbi.nlm.nih.gov/protein/P09488
  - Sometimes, you will need more information to get the desired format (e.g. FASTA)
    http://www.ncbi.nlm.nih.gov/protein/121735?report=fasta
- curl and wget allow you to pull a web page into a file from the command line:
  curl http://uniprot.org/uniprot/P09488.fasta > p09488.fasta
- Sometimes this is what you need; other times more work is required

Finding a URL (www.uniprot.org)



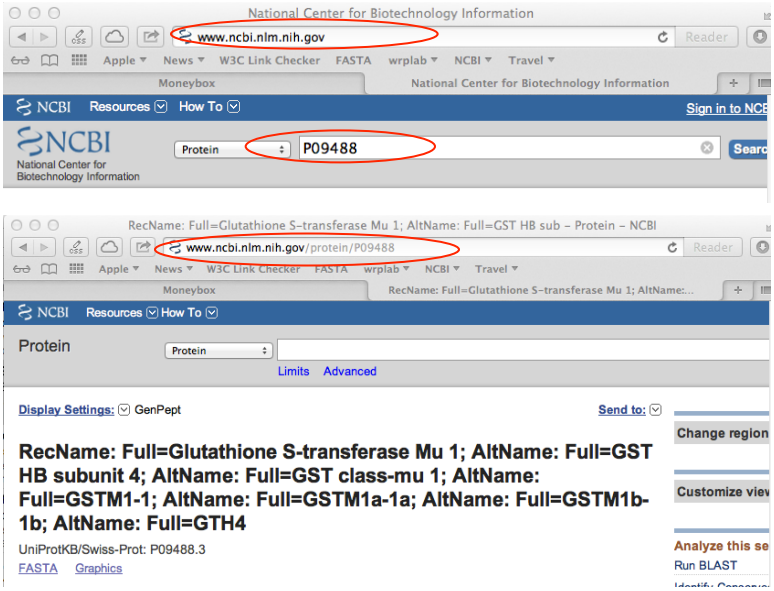
Finding a URL to download (uniprot)




```
>sp|P09488|GSTM1_HUMAN Glutathione S-transferase Mu 1 OS=Homo sapiens GN=GSTM1 PE=1 SV=3
MPMILGYWDIRGLAHAIRLLLEYTDSSYEEKKYTMGDAPDYDRSQWLNEKFKLGLDFPNL
PYLIDGAHKITQSNAILCYIARKHNLCGETEEEKIRVDILENQTMDNHMQLGMICYNPEF
EKLKPKYLEELPEKLKLYSEFLGKRPWFAGNKITFVDFLVYDVLDLHRIFEPKCLDAFPN
LKDFISRFEGLEKISAYMKSSRFLPRPVFSKMAVWGNK
```

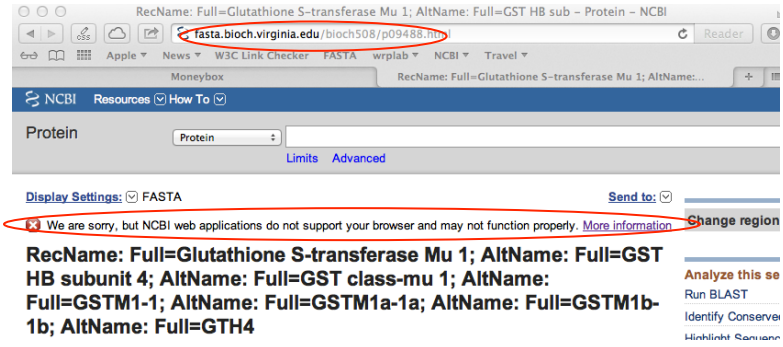
```
curl http://www.uniprot.org/uniprot/P09488.fasta
```

Finding a URL to download (NCBI)



Finding a URL to download (NCBI)

```
curl http://www.ncbi.nlm.nih.gov/protein/121735?report=fasta
```



NCBI e-utilities

- The NCBI does not allow their web server to be used for large-scale, automated downloads (unlike Uniprot)
  www.ncbi.nlm.nih.gov/guide/howto/dwn-records/
- NCBI provides e-utilities (esearch.cgi, efetch.cgi) for programmatic access to ALL NCBI databases (proteins, DNA, also PubMed)
  www.ncbi.nlm.nih.gov/books/NBK25500/
- e-utilities need an Entrez GI number (e.g. 121735 for P09488.1)
- With a GI number, downloading a fasta file is easy (possible):

```
curl 'http://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?
db=protein&id=121735&rettype=fasta&retmode=text'
```

  Quotes are required to protect '&' and '?' from shell

- (to keep things simple, use UniProt)

Homework

From the tabular blastp search you did earlier (-outfmt 6) against swissprot:

1. Edit the file to save 10 lines with E()-values > 0.1
2. isolate the library (subject) accession information for each of the 10 lines in the edited file, and save the accession in a new file
3. For each accession, split it into its component parts (hit 'man cut' to see how to change the delimiter).
   - Save the gi numbers for use with NCBI in one file
   - Isolate the accessions (P12345.3) in another file.
   - For this second file, isolate only the accessions without the version information.
4. For of the 10 gi numbers or accessions
   - Use the gi numbers to get the sequences from the NCBI
   - Use the protein accessions to get the sequences from UniProt
5. For each of the 10 UniProt accessions, run another blastp search against swissprot, saving the output to accession.bp

   (e.g. `blastp –db swissprot –query P12345.aa > P12345.bp`)