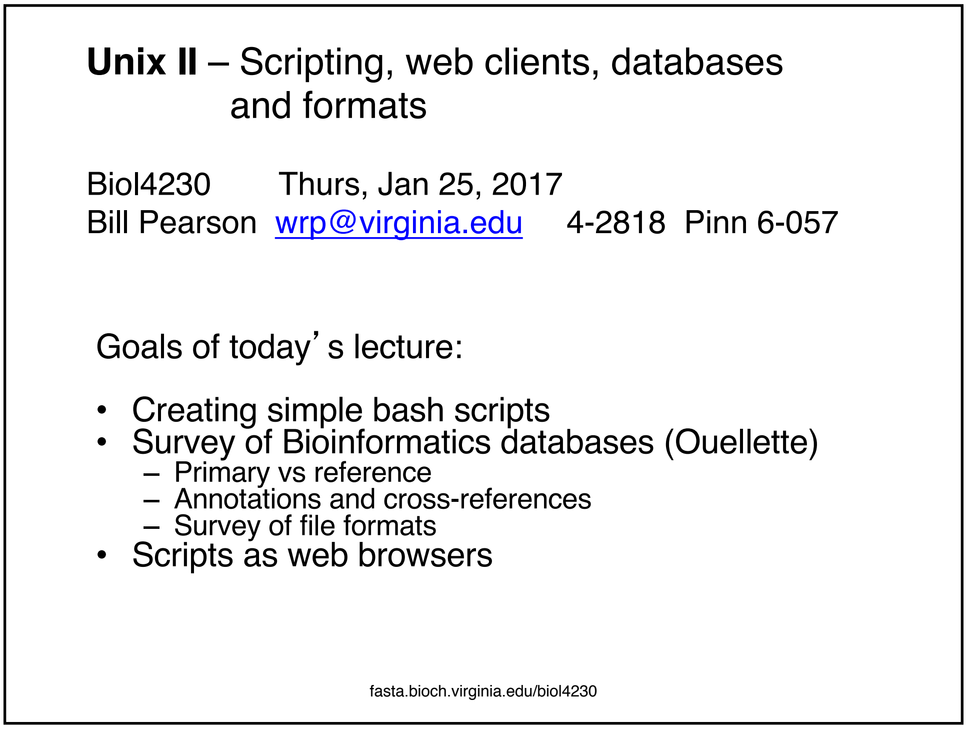


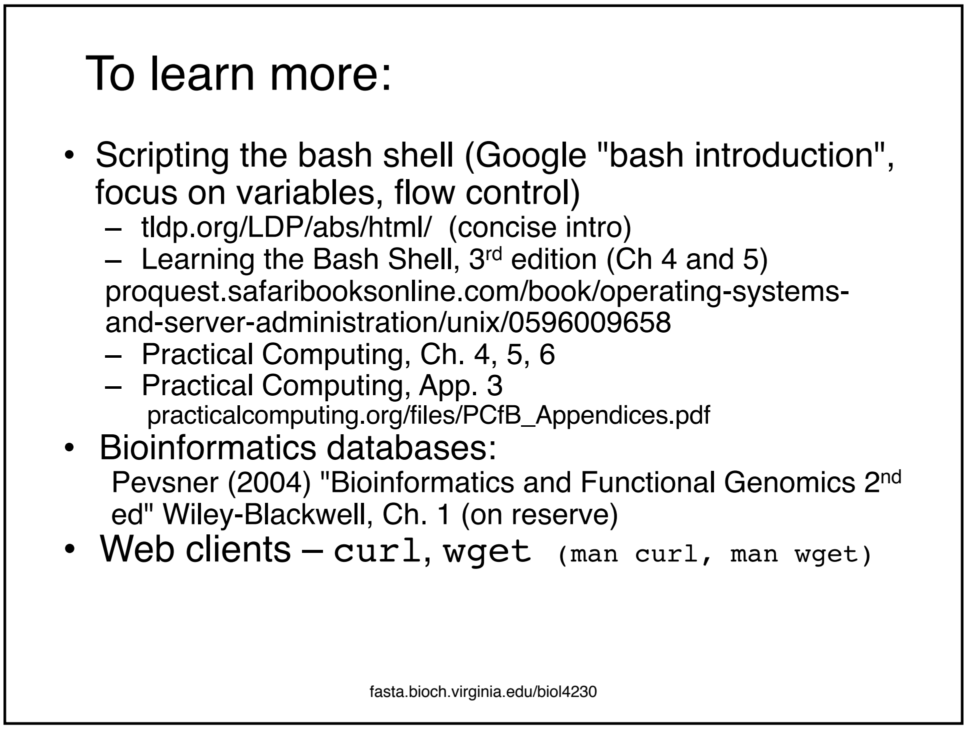
## **Unix II** – Scripting, web clients, databases and formats

Biol4230 Thurs, Jan 25, 2017
Bill Pearson wrp@virginia.edu 4-2818 Pinn 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

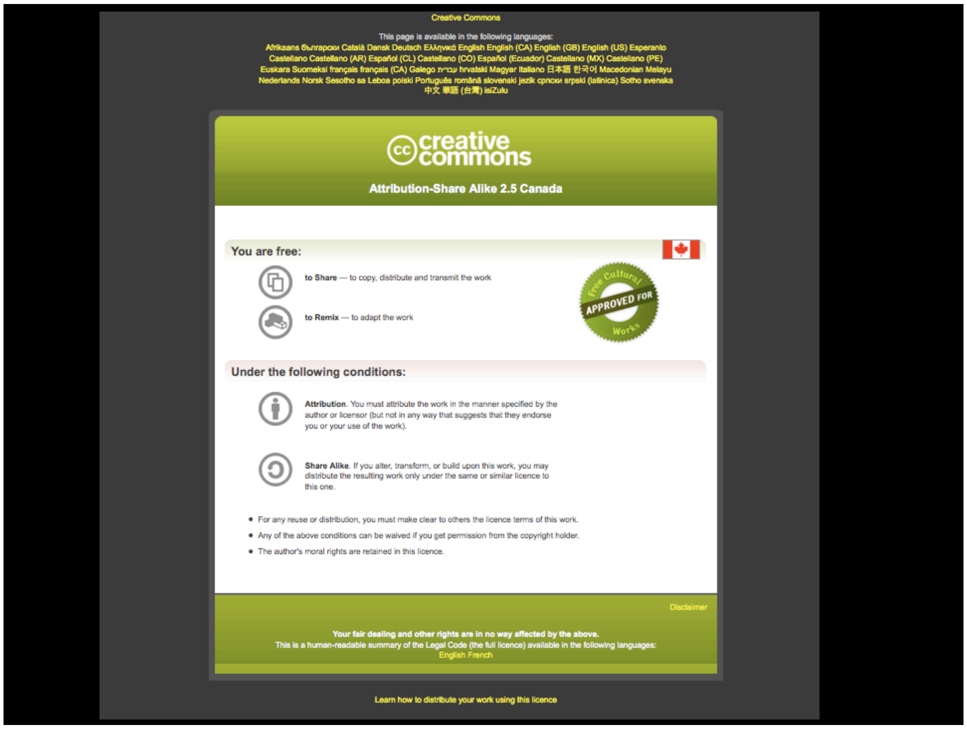
fasta.bioch.virginia.edu/biol4230



## To learn more:

- 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
  - Practical Computing, Ch. 4, 5, 6
  - Practical Computing, App. 3 practicalcomputing.org/files/PCfB_Appendices.pdf
- Bioinformatics databases: Pevsner (2004) "Bioinformatics and Functional Genomics 2nd ed" Wiley-Blackwell, Ch. 1 (on reserve)
- Web clients – `curl`, `wget` `(man curl, man wget)`

fasta.bioch.virginia.edu/biol4230



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

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;

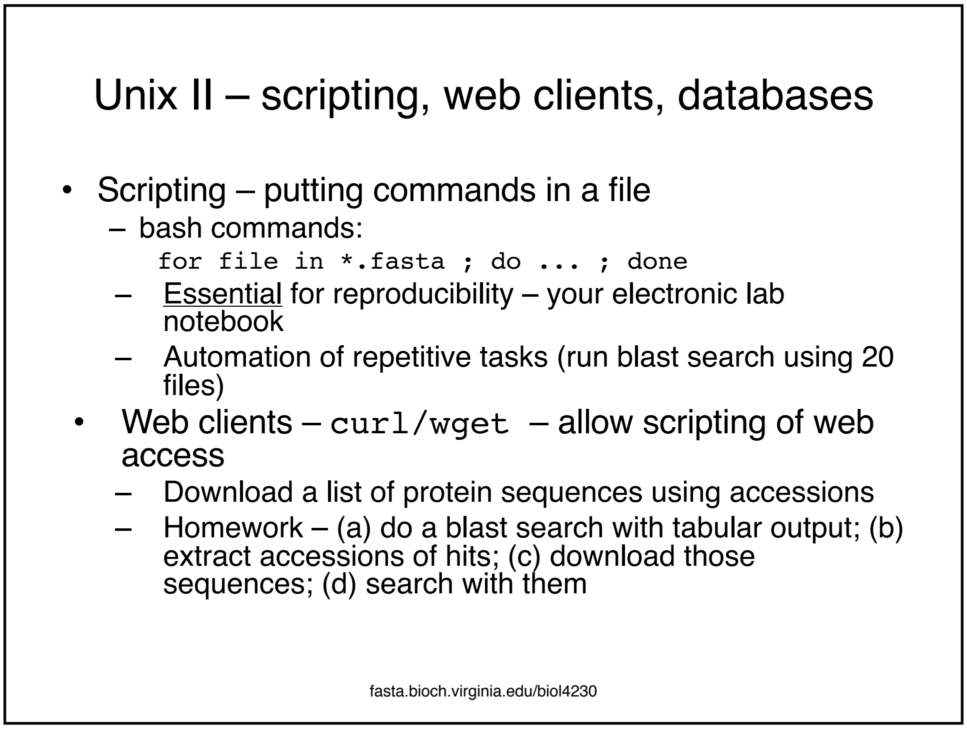
Photograph, film, or broadcast;

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.

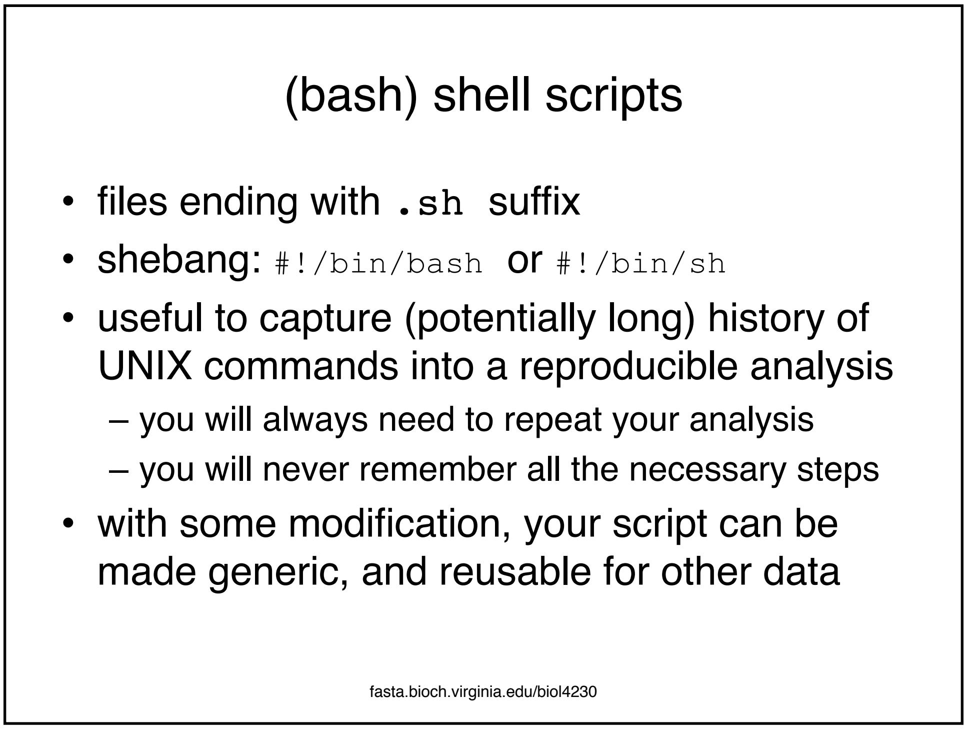
Slide Concept by Cameron Neylon, who has waived all copyright and related or neighbouring rights. This slide only ccZero.
Social Media Icons adapted with permission from originals by Christopher Ross. Original images are available under GPL at;
http://www.thisismyurl.com/free-downloads/15-free-speech-bubble-icons-for-popular-websites



## 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

fasta.bioch.virginia.edu/biol4230



## (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

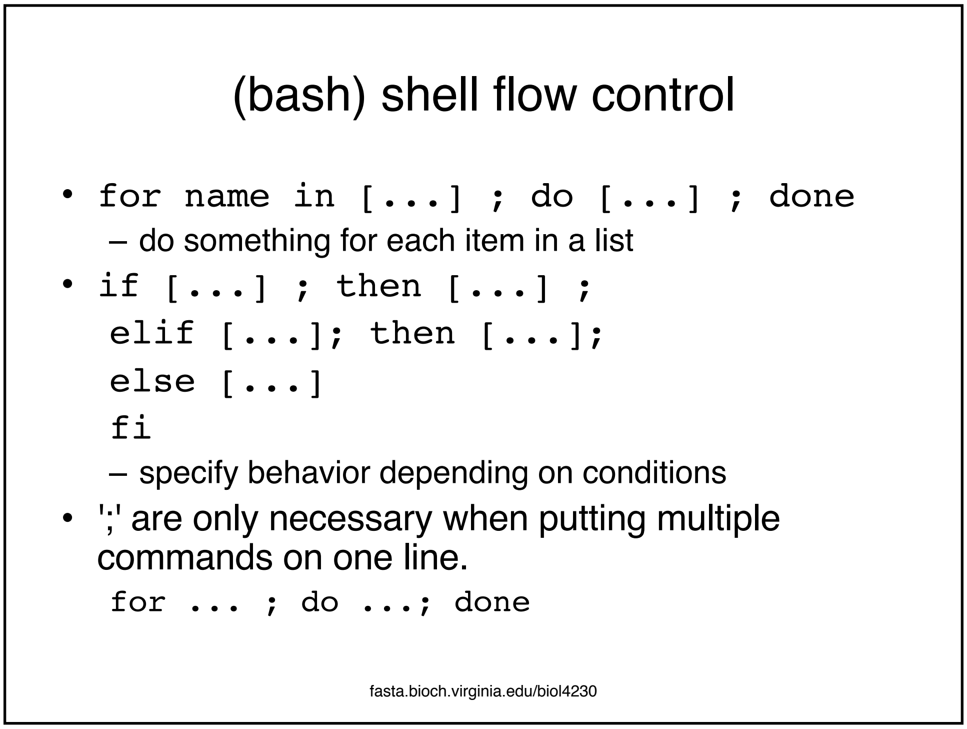
fasta.bioch.virginia.edu/biol4230

# 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 `$NAME` (env) / `$name` (shell)
  - 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 '`$`' on the right side.
  - no spaces around the '='
  ```
  new_files=$files
  ```
- $SHELL variables are transient; to make them permanent, use:
  ```
  export PATH=$PATH:/seqprg/bin
  ```



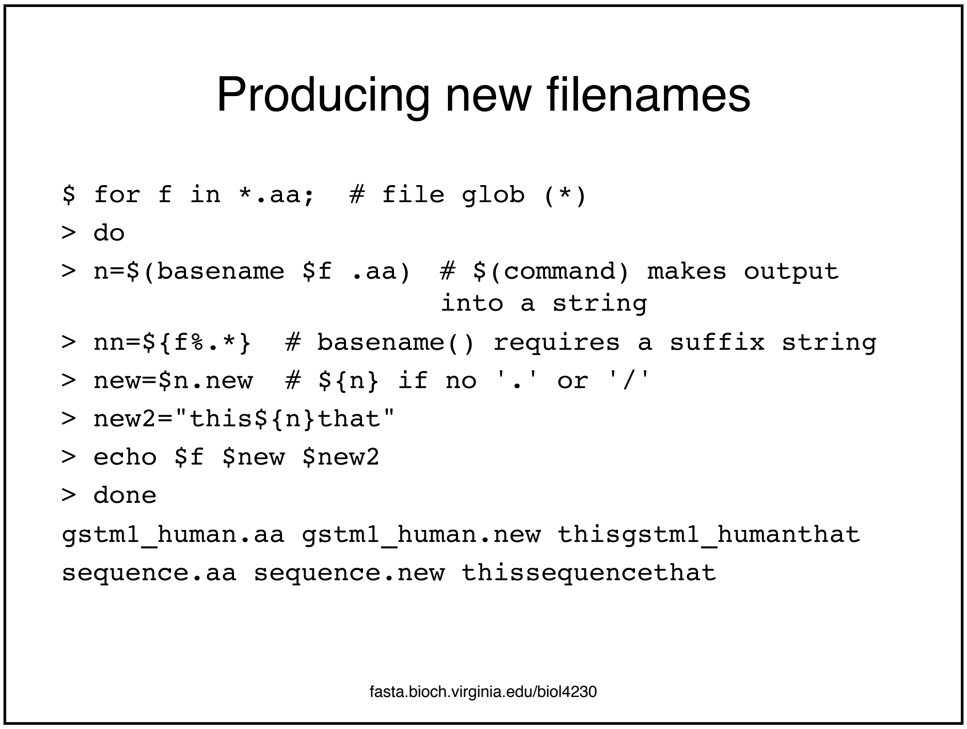
# (bash) shell flow control

- `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
```

fasta.bioch.virginia.edu/biol4230



# Producing new filenames

```
$ for f in *.aa;  # file glob (*)
> do
> n=$(basename $f .aa)  # $(command) makes output
                        into a string
> nn=${f%.*}  # basename() requires a suffix 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
```

fasta.bioch.virginia.edu/biol4230

# Extracting parts of lines: `cut`

```
# do a blastp search:
$ blastp –outfmt 6 –query atp6_human.aa –d /slib2/bl_dbs/pir1 > atp6.bl_out

# look at first three lines
$ head –n 3 atp6.bl_out
sp|P00846|ATP6_HUMAN    P00846  100.000 226     0       0       1       226     1       226     2.00e-157       434
sp|P00846|ATP6_HUMAN    P00847  77.876  226     50      0       1       226     1       226     3.58e-124       349
sp|P00846|ATP6_HUMAN    P00848  75.664  226     55      0       1       226     1       226     5.66e-112       318

#qsid            ssid perc  alen  mism gaps  qstart qend sstart send evalue    bits

# extract only the ssid column (field)
$ cut -f 2 atp6.bl_out | head -n 3
P00846
P00847
P00848
# change field delimiter with -d " ", -d "|", etc.
```

CSH
Meetings & Courses Program
Cold Spring Harbor Laboratory
Home
Meetings
Courses
How To Apply
Selection Process and Stipends
Travel
General Information

COMPUTATIONAL & COMPARATIVE GENOMICS
October 31 - November 6, 2012
Application Deadline: July 15, 2012

Instructors:
William Pearson, *University of Virginia, Charlottesville*
Lisa Stubbs, *University of Illinois, Urbana*

# COMPUTATIONAL & COMPARATIVE GENOMICS: Understanding and Using Biological Databases

November 30th, 2012

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

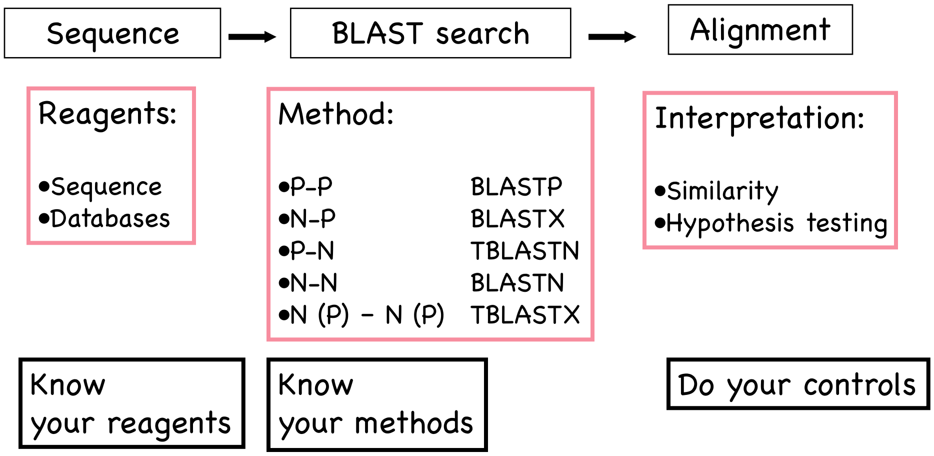
- Associate Director, Informatics and Biocomputing, Ontario Institute for Cancer Research, Toronto, ON
- Associate Professor, Department of Cell and Systems Biology, University of Toronto, Toronto, ON.

## 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:



Sequence → BLAST search → Alignment

Reagents:

- Sequence
- Databases

Method:

| | |
|---|---|
| •P-P | BLASTP |
| •N-P | BLASTX |
| •P-N | TBLASTN |
| •N-N | BLASTN |
| •N (P) – N (P) | TBLASTX |

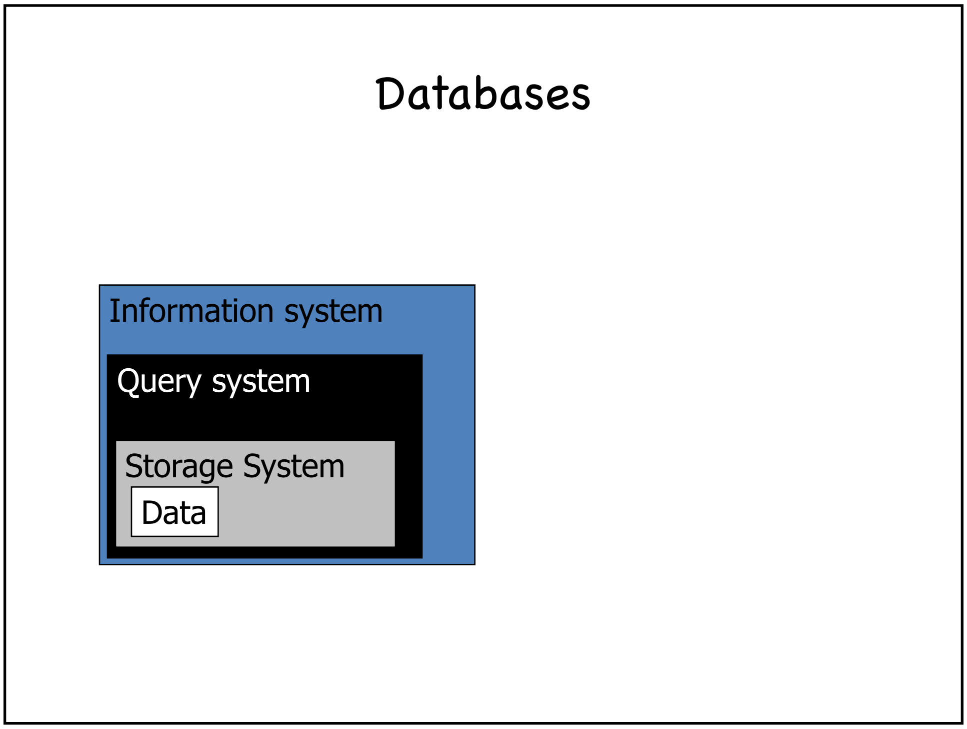
Interpretation:

- Similarity
- Hypothesis testing

Know your reagents

Know your methods

Do your controls



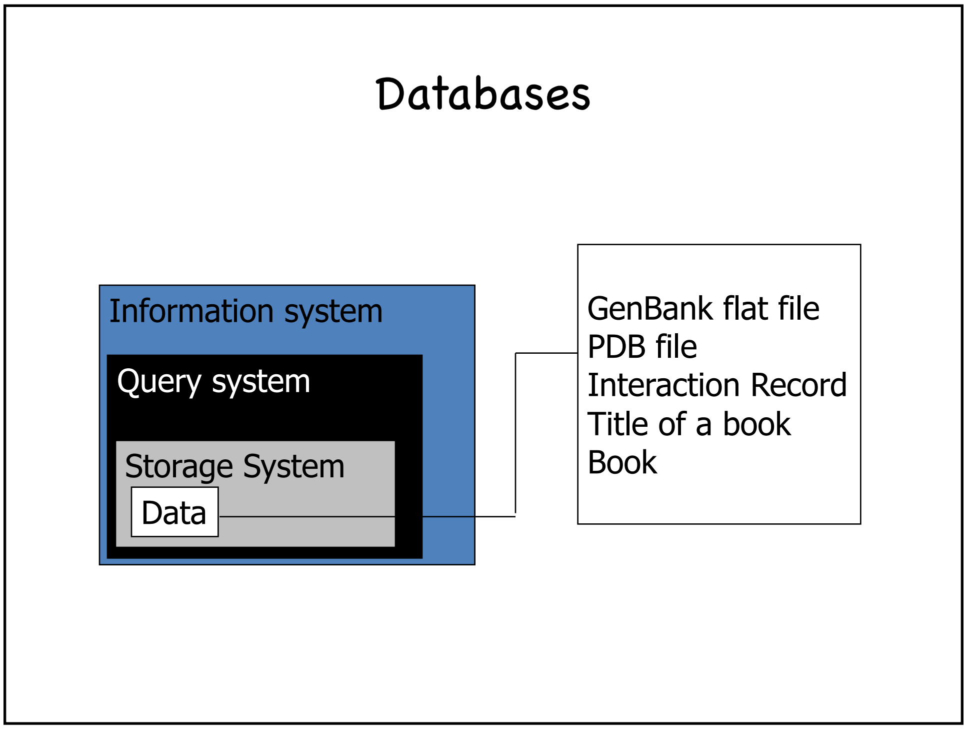
## Databases

Information system

Query system

Storage System

Data



## Databases

Information system

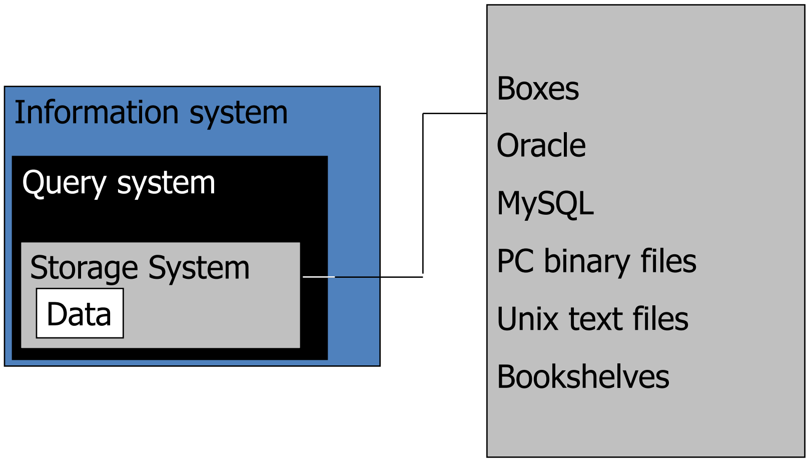
Query system

Storage System

Data

GenBank flat file
PDB file
Interaction Record
Title of a book
Book

# Databases



Information system

Query system

Storage System

Data

Boxes

Oracle

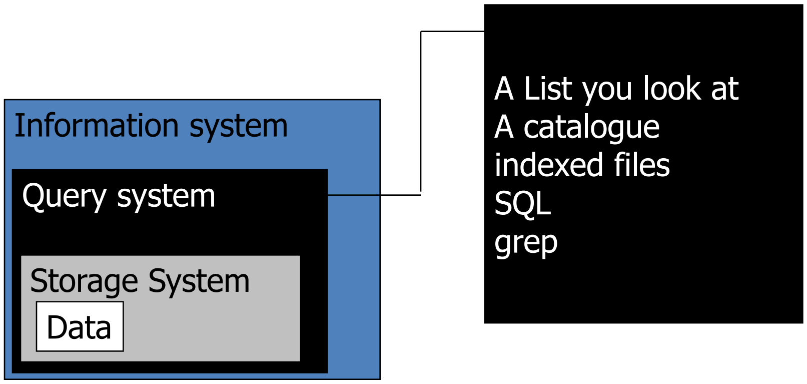
MySQL

PC binary files

Unix text files

Bookshelves

# Databases



Information system

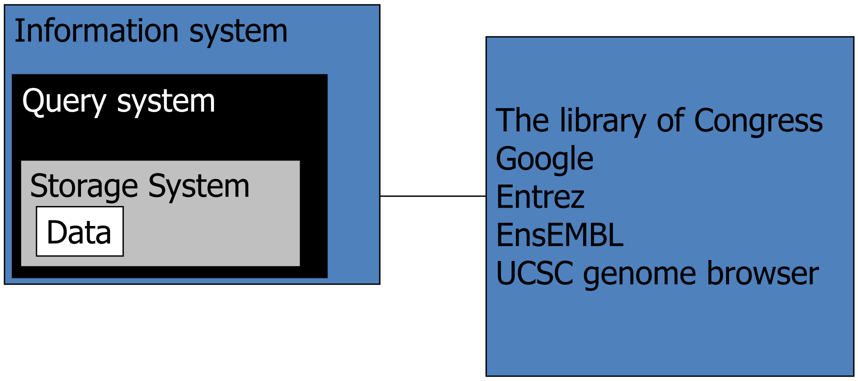
Query system

Storage System

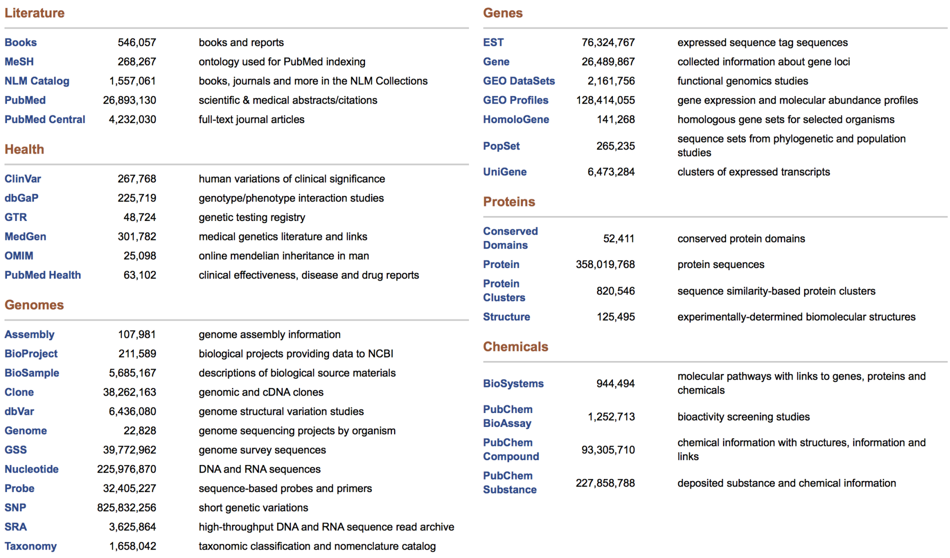
Data

A List you look at
A catalogue
indexed files
SQL
grep

# Databases



## www.ncbi.nlm.nih.gov/gquery/?term=all[filter]



**Literature**

| | | |
|---|---|---|
| Books | 546,057 | books and reports |
| MeSH | 268,267 | ontology used for PubMed indexing |
| NLM Catalog | 1,557,061 | books, journals and more in the NLM Collections |
| PubMed | 26,893,130 | scientific & medical abstracts/citations |
| PubMed Central | 4,232,030 | full-text journal articles |

**Health**

| | | |
|---|---|---|
| ClinVar | 267,768 | human variations of clinical significance |
| dbGaP | 225,719 | genotype/phenotype interaction studies |
| GTR | 48,724 | genetic testing registry |
| MedGen | 301,782 | medical genetics literature and links |
| OMIM | 25,098 | online mendelian inheritance in man |
| PubMed Health | 63,102 | clinical effectiveness, disease and drug reports |

**Genomes**

| | | |
|---|---|---|
| Assembly | 107,981 | genome assembly information |
| BioProject | 211,589 | biological projects providing data to NCBI |
| BioSample | 5,685,167 | descriptions of biological source materials |
| Clone | 38,262,163 | genomic and cDNA clones |
| dbVar | 6,436,080 | genome structural variation studies |
| Genome | 22,828 | genome sequencing projects by organism |
| GSS | 39,772,962 | genome survey sequences |
| Nucleotide | 225,976,870 | DNA and RNA sequences |
| Probe | 32,405,227 | sequence-based probes and primers |
| SNP | 825,832,256 | short genetic variations |
| SRA | 3,625,864 | high-throughput DNA and RNA sequence read archive |
| Taxonomy | 1,658,042 | taxonomic classification and nomenclature catalog |

**Genes**

| | | |
|---|---|---|
| EST | 76,324,767 | expressed sequence tag sequences |
| Gene | 26,489,867 | collected information about gene loci |
| GEO DataSets | 2,161,756 | functional genomics studies |
| GEO Profiles | 128,414,055 | gene expression and molecular abundance profiles |
| HomoloGene | 141,268 | homologous gene sets for selected organisms |
| PopSet | 265,235 | sequence sets from phylogenetic and population studies |
| UniGene | 6,473,284 | clusters of expressed transcripts |

**Proteins**

| | | |
|---|---|---|
| Conserved Domains | 52,411 | conserved protein domains |
| Protein | 358,019,768 | protein sequences |
| Protein Clusters | 820,546 | sequence similarity-based protein clusters |
| Structure | 125,495 | experimentally-determined biomolecular structures |

**Chemicals**

| | | |
|---|---|---|
| BioSystems | 944,494 | molecular pathways with links to genes, proteins and chemicals |
| PubChem BioAssay | 1,252,713 | bioactivity screening studies |
| PubChem Compound | 93,305,710 | chemical information with structures, information and links |
| PubChem Substance | 227,858,788 | deposited substance and chemical information |

http://www.ncbi.nlm.nih.gov/
All [filter] Jan, 2017

## 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
KEYWORDS    .
SOURCE      Prunus salicina
  ORGANISM  Prunus salicina
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; Gunneridae;
            Pentapetalae; rosids; fabids; Rosales; Rosaceae; Maloideae;
            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             128..1249
                     /note="MAPK 1"
                     /codon_start=1
                     /product="mitogen-activated protein kinase 1"
                     /protein_id="AEQ28763.1"
                     /translation="MESSSASAGDHNIKGVPAHGGRYVQYNVYGNFFEVSRKYVPPIR
                     PVGRGAYGIVCAAVNAETREEVAIKKIGNAFDNRIDAKRTLREIKLLRHMDHENVIAI
                     KDIIRPPQKENFNDVYIVYELMDTDLHQIIRSNQPLNDDHCRYFLYQLLRGLKYVHSA
                     NVLHRDLKPSNLLMNANCDLKIGDFGLARTTSETDFMTEYVVTRWYRAPELLLNCSEY
                     TAAIDIWSVGCILGEIMTRRPLFPGKDYVHQLRLITELLGSPDDSSLGFLRSDNARRY
                     VRQLPQYPKQSFSAGFPNMSPGAVDLLEKMLVFDPNRRITVDEALCHPYLAPLHDINE
                     EPVCPMPFNFDFEQPSFTEENIKELIWRESVKFNPDPIH"
ORIGIN
        1 ccattacggc ctagttacgg gggatcatta gtgctctcta tagctccttc tctacaagtc
       61 ttagcctttt agcaccgaaa tcctaactgg ttcttgaccc atttcggatt cggatccctg
      121 gttaactatg gaatccagct ctgcttcagc aggtgatcac aatatcaaag gggtacctgc
      181 ccacggtgga cgctatgttc agtacaatgt gtacggtaac ttctttgagg tttctaggaa
      241 gtacgtccct cccataaggc ccgtagggag aggtgcttat ggtattgttt gtgctgctgt
      301 gaatgctgag actcgtgagg aggttgccat taagaagatt ggtaatgcat ttgacaacag
      361 aattgatgcc aagaggactt tacgagaaat taaacttctt cggcacatgg atcatgaaaa
      421 tgttattgcc atcaaagaca tcatacggcc tccacagaag gagaacttca atgatgtcta
...
     1621 aaggtaggtc tgaatataaa cggttgcctt ttttccaaaa gaaaaaaaaa aaaaaaaaaa
     1681 aaaaaaaaaa aaaaaaaaaa aaaa
//
```

Header
- •Title
- •Taxonomy
- •Citation

Features (AA seq)

DNA Sequence

# FASTA

NCBI

```
>P03069.1 RecName: Full=General control protein GCN4; ...
MSEYQPSLFALNPMGFSPLDGSKSTNENVSASTSTAKPMVGQLIFDKFIKTEEDPI
IKQDTPSNLDFDFALPQTATAPDAKTVLPIPELDDAVVESFFSSSTDSTPMFEYEN
LEDNSKEWTSLFDNDIPVTTDDVSLADKAIESTEEVSLVPSNLEVSTTSFLPTPVL
EDAKLTQTRKVKKPNSVVKKSHHVGKDDESRLDHLGVVAYNRKQRSIPLSPIVPES
SDPAALKRARNTEAARRSRARKLQRMKQLEDKVEELLSKNYHLENEVARLKKLVGE
R
```

uniprot.org

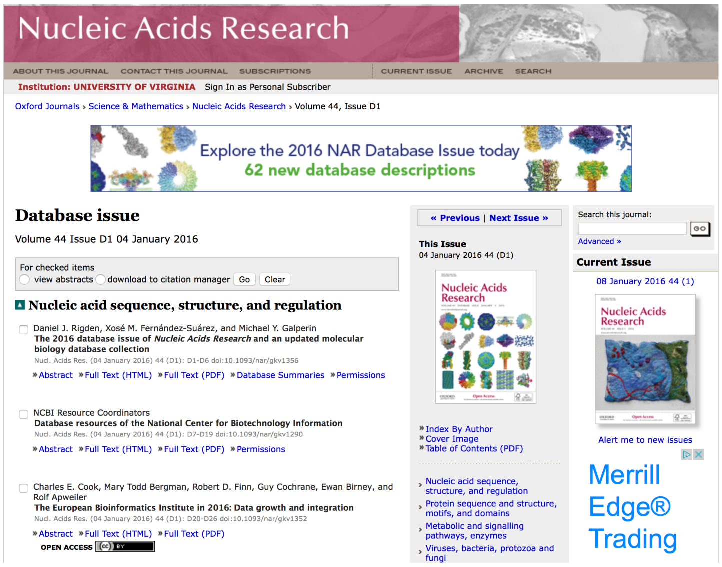
```
>sp|P03069|GCN4_YEAST General control protein GCN4 ... GN=GCN4 PE=1 SV=1
MSEYQPSLFALNPMGFSPLDGSKSTNENVSASTSTAKPMVGQLIFDKFIKTEEDPIIKQD
TPSNLDFDFALPQTATAPDAKTVLPIPELDDAVVESFFSSSTDSTPMFEYENLEDNSKEW
TSLFDNDIPVTTDDVSLADKAIESTEEVSLVPSNLEVSTTSFLPTPVLEDAKLTQTRKVK
KPNSVVKKSHHVGKDDESRLDHLGVVAYNRKQRSIPLSPIVPESSDPAALKRARNTEAAR
RSRARKLQRMKQLEDKVEELLSKNYHLENEVARLKKLVGER
```

# Databases

- Primary (archival)
  - GenBank/EMBL/DDBJ
  - UniProt
  - PDB
  - Medline (PubMed)
  - Intact

- Secondary (curated)
  - RefSeq
  - Taxon
  - UniProt
  - OMIM
  - SGD
  - Biosamples/Bioprojects

https://nar.oxfordjournals.org/content/45/D1.toc January 2017



Nucleic Acids Research

ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS CURRENT ISSUE ARCHIVE SEARCH

Institution: UNIVERSITY OF VIRGINIA Sign In as Personal Subscriber

Oxford Journals › Science & Mathematics › Nucleic Acids Research › Volume 44, Issue D1

Explore the 2016 NAR Database Issue today
62 new database descriptions

## Database issue

Volume 44 Issue D1 04 January 2016

For checked items
○ view abstracts ○ download to citation manager Go Clear

### Nucleic acid sequence, structure, and regulation

□ Daniel J. Rigden, Xosé M. Fernández-Suárez, and Michael Y. Galperin
**The 2016 database issue of *Nucleic Acids Research* and an updated molecular biology database collection**
Nucl. Acids Res. (04 January 2016) 44 (D1): D1-D6 doi:10.1093/nar/gkv1356
» Abstract » Full Text (HTML) » Full Text (PDF) » Database Summaries » Permissions

□ NCBI Resource Coordinators
**Database resources of the National Center for Biotechnology Information**
Nucl. Acids Res. (04 January 2016) 44 (D1): D7-D19 doi:10.1093/nar/gkv1290
» Abstract » Full Text (HTML) » Full Text (PDF) » Permissions

□ Charles E. Cook, Mary Todd Bergman, Robert D. Finn, Guy Cochrane, Ewan Birney, and Rolf Apweiler
**The European Bioinformatics Institute in 2016: Data growth and integration**
Nucl. Acids Res. (04 January 2016) 44 (D1): D20-D26 doi:10.1093/nar/gkv1352
» Abstract » Full Text (HTML) » Full Text (PDF)
OPEN ACCESS (cc) BY

« Previous | Next Issue »

**This Issue**
04 January 2016 44 (D1)

» Index By Author
» Cover Image
» Table of Contents (PDF)

› Nucleic acid sequence, structure, and regulation
› Protein sequence and structure, motifs, and domains
› Metabolic and signalling pathways, enzymes
› Viruses, bacteria, protozoa and fungi

Search this journal:
GO
Advanced »

**Current Issue**
08 January 2016 44 (1)

Alert me to new issues

# Sequence Databases

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

# 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 )))

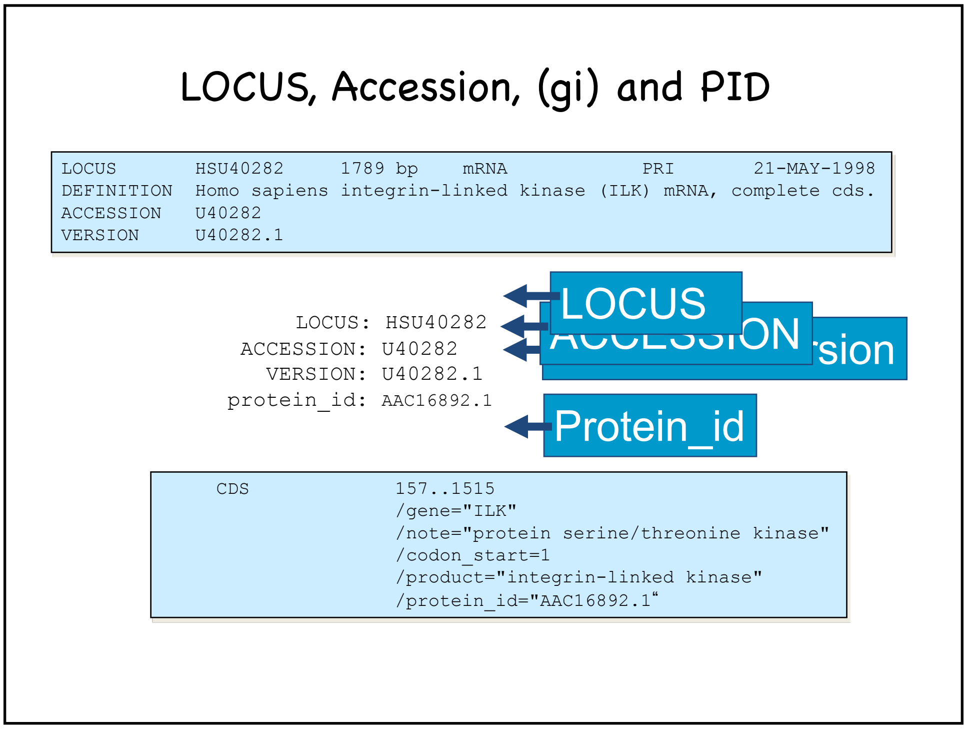
## LOCUS, Accession, NID and protein_id

**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.

**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
```

```
     LOCUS: HSU40282
 ACCESSION: U40282
   VERSION: U40282.1
protein_id: AAC16892.1
```

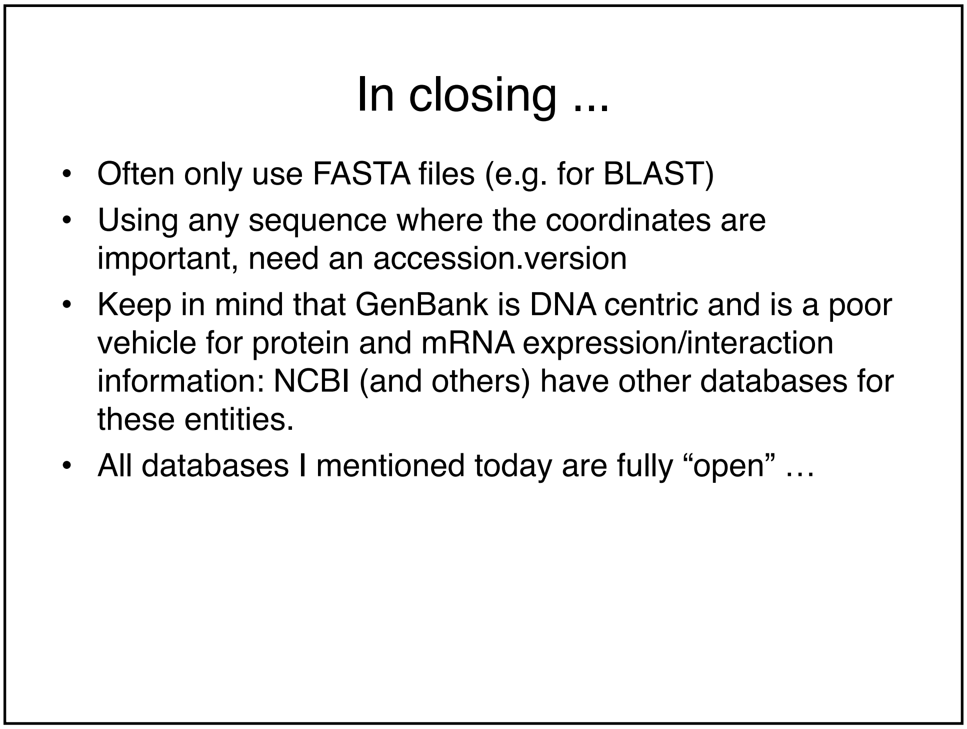
LOCUS

ACCESSION

sion

Protein_id

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



## In closing ...

- Often only use FASTA files (e.g. for BLAST)
- Using any sequence where the coordinates are important, need an accession.version
- 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" …

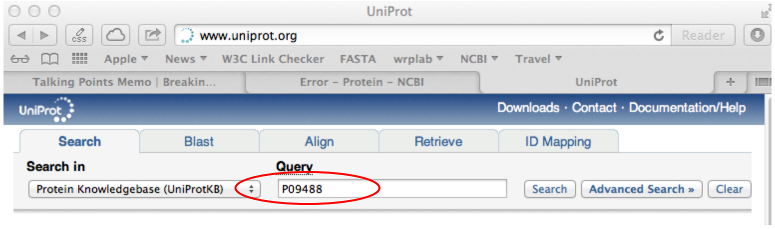
# Scripting from the WWW: `wget/curl`

- Most bioinformatics 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

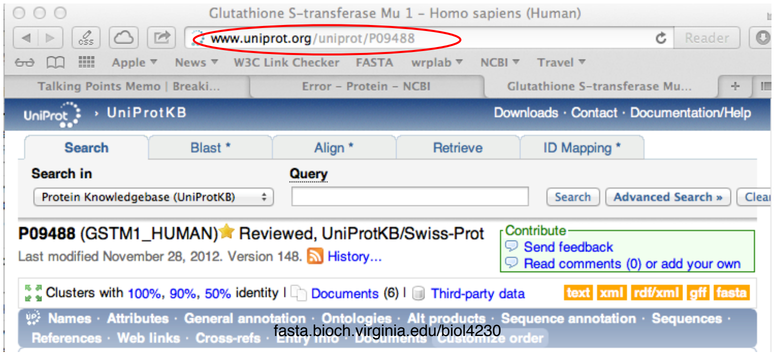
# Finding a URL (www.uniprot.org)



www.uniprot.org

Search in: Protein Knowledgebase (UniProtKB)

Query: P09488


Glutathione S–transferase Mu 1 – Homo sapiens (Human)
Reader
Apple
News
W3C Link Checker
FASTA
wrplab
NCBI
Travel
Talking Points Memo | Breaki...
Error – Protein – NCBI
Glutathione S–transferase Mu...
UniProt
UniProtKB
Downloads · Contact · Documentation/Help
Search
Blast *
Align *
Retrieve
ID Mapping *
Search
Advanced Search »
Clear
Contribute
Send feedback
Read comments (0) or add your own
text
xml
rdf/xml
gff
fasta
Names · Attributes · General annotation · Ontologies · Alt products · Sequence annotation · Sequences ·
References · Web links · Cross-refs · Entry info · Documents · Customize order

www.uniprot.org/uniprot/P09488

Search in: Protein Knowledgebase (UniProtKB)

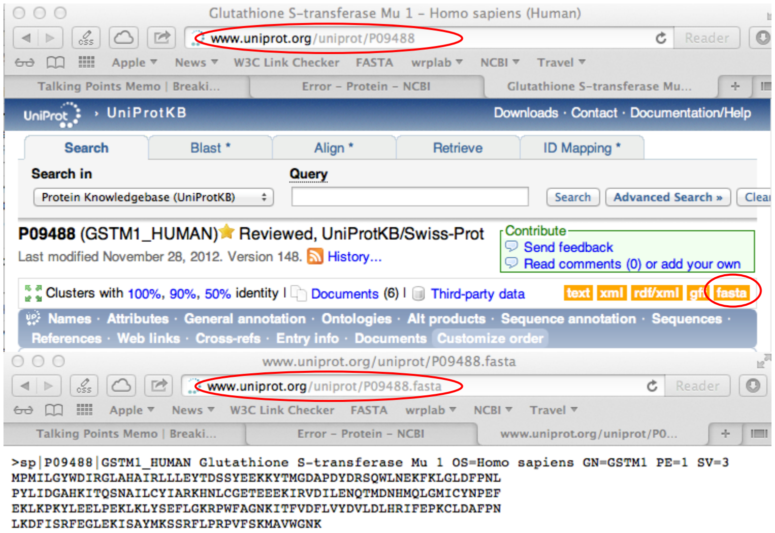
Query:

P09488 (GSTM1_HUMAN) Reviewed, UniProtKB/Swiss-Prot

Last modified November 28, 2012. Version 148. History...

Clusters with 100%, 90%, 50% identity | Documents (6) | Third-party data

# Finding a URL to download (uniprot)


Glutathione S-transferase Mu 1 – Homo sapiens (Human)
www.uniprot.org/uniprot/P09488
Reader
Apple News W3C Link Checker FASTA wrplab NCBI Travel
Talking Points Memo | Breaki... Error – Protein – NCBI Glutathione S-transferase Mu...
UniProt UniProtKB
Downloads · Contact · Documentation/Help
Search Blast * Align * Retrieve ID Mapping *
Search in
Query
Protein Knowledgebase (UniProtKB)
Search Advanced Search » Clear
Contribute
Send feedback
Read comments (0) or add your own
Clusters with 100%, 90%, 50% identity | Documents (6) | Third-party data
text xml rdf/xml gff fasta
Names · Attributes · General annotation · Ontologies · Alt products · Sequence annotation · Sequences · References · Web links · Cross-refs · Entry info · Documents Customize order
www.uniprot.org/uniprot/P09488.fasta
www.uniprot.org/uniprot/P09488.fasta
Talking Points Memo | Breaki... Error – Protein – NCBI www.uniprot.org/uniprot/P0...

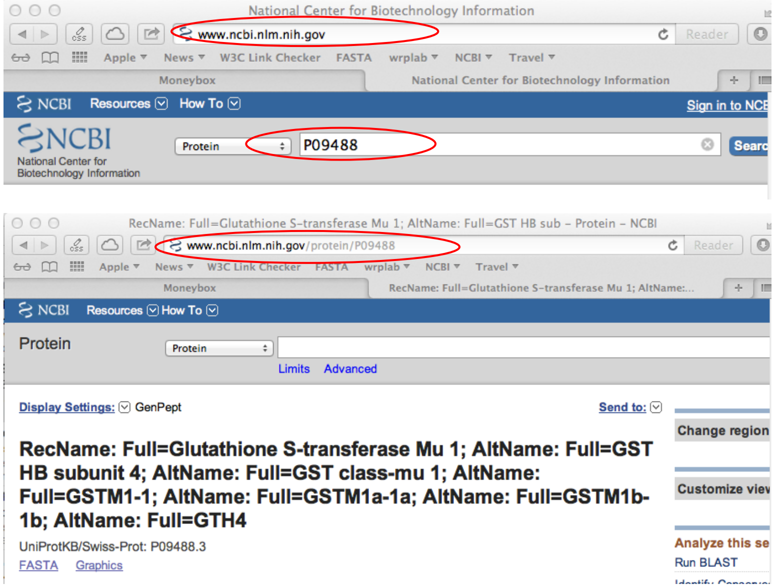
P09488 (GSTM1_HUMAN) Reviewed, UniProtKB/Swiss-Prot
Last modified November 28, 2012. Version 148. History...

```
>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
```

fasta.bioch.virginia.edu/biol4230

# Finding a URL to download (NCBI)


National Center for Biotechnology Information
www.ncbi.nlm.nih.gov
Reader
Apple News W3C Link Checker FASTA wrplab NCBI Travel
Moneybox National Center for Biotechnology Information
NCBI Resources How To Sign in to NCBI
NCBI National Center for Biotechnology Information
Protein
P09488
Search
RecName: Full=Glutathione S-transferase Mu 1; AltName: Full=GST HB sub – Protein – NCBI
www.ncbi.nlm.nih.gov/protein/P09488
Moneybox RecName: Full=Glutathione S-transferase Mu 1; AltName:...
NCBI Resources How To
Protein
Protein
Limits Advanced
Display Settings: GenPept
Send to:
Change region
Customize view
Analyze this se
Run BLAST

RecName: Full=Glutathione S-transferase Mu 1; AltName: Full=GST HB subunit 4; AltName: Full=GST class-mu 1; AltName: Full=GSTM1-1; AltName: Full=GSTM1a-1a; AltName: Full=GSTM1b-1b; AltName: Full=GTH4

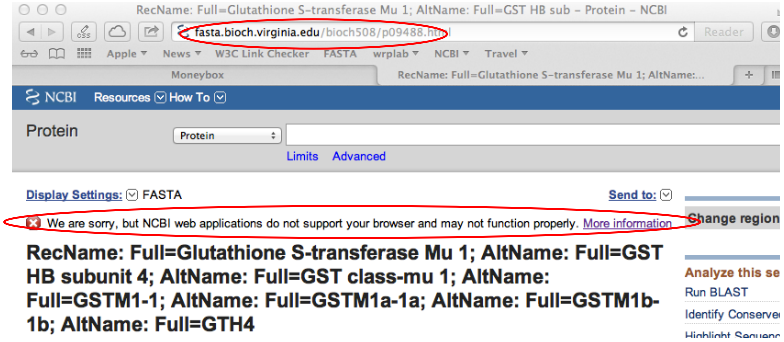
UniProtKB/Swiss-Prot: P09488.3

FASTA Graphics

fasta.bioch.virginia.edu/biol4230

# Finding a URL to download (NCBI)

```
curl http://www.ncbi.nlm.nih.gov/protein/P09488?report=fasta
```



We are sorry, but NCBI web applications do not support your browser and may not function properly. More information

**RecName: Full=Glutathione S-transferase Mu 1; AltName: Full=GST HB subunit 4; AltName: Full=GST class-mu 1; AltName: Full=GSTM1-1; AltName: Full=GSTM1a-1a; AltName: Full=GSTM1b-1b; AltName: Full=GTH4**

# 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/

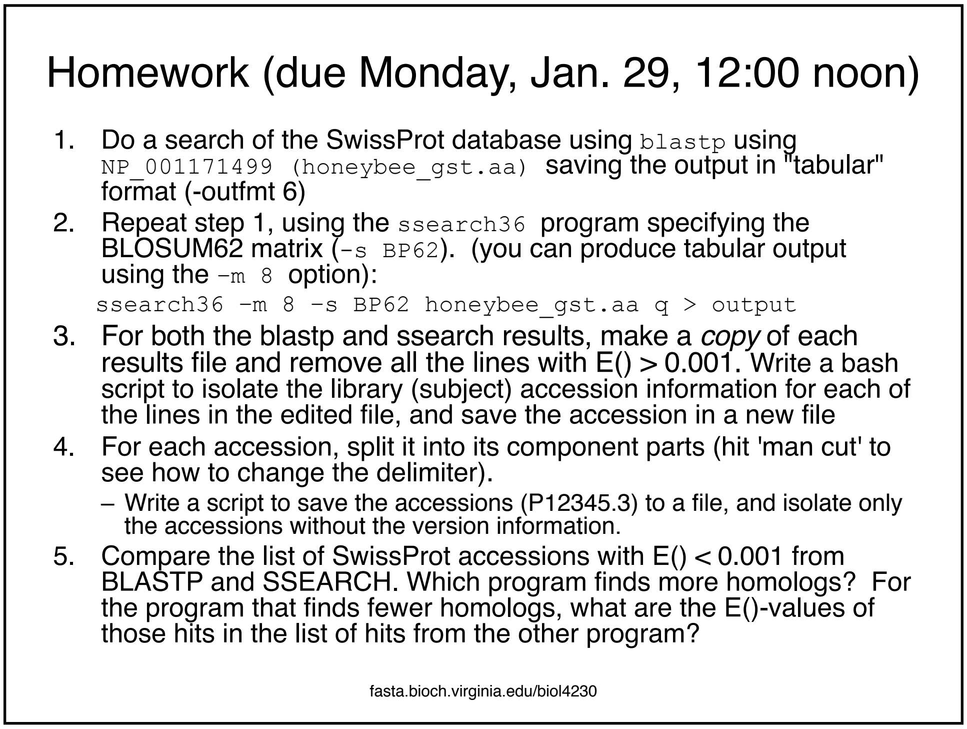
- NCBI provides e-utilities (esearch.cgi, efetch.cgi) for programmatic access to ALL NCBI databases (proteins, DNA, also PubMed)

  www.ncbi.nlm.nih.gov/books/NBK25500/ (this document is currenty out of date because it still users GI numbers)

  In 2017, NCBI also uses accessions for downloads, so downloading a fasta file is easy:

  ```
  curl
  'https://eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.f
  cgi?db=protein&id=P09488&rettype=fasta&retmode=text'
  ```

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

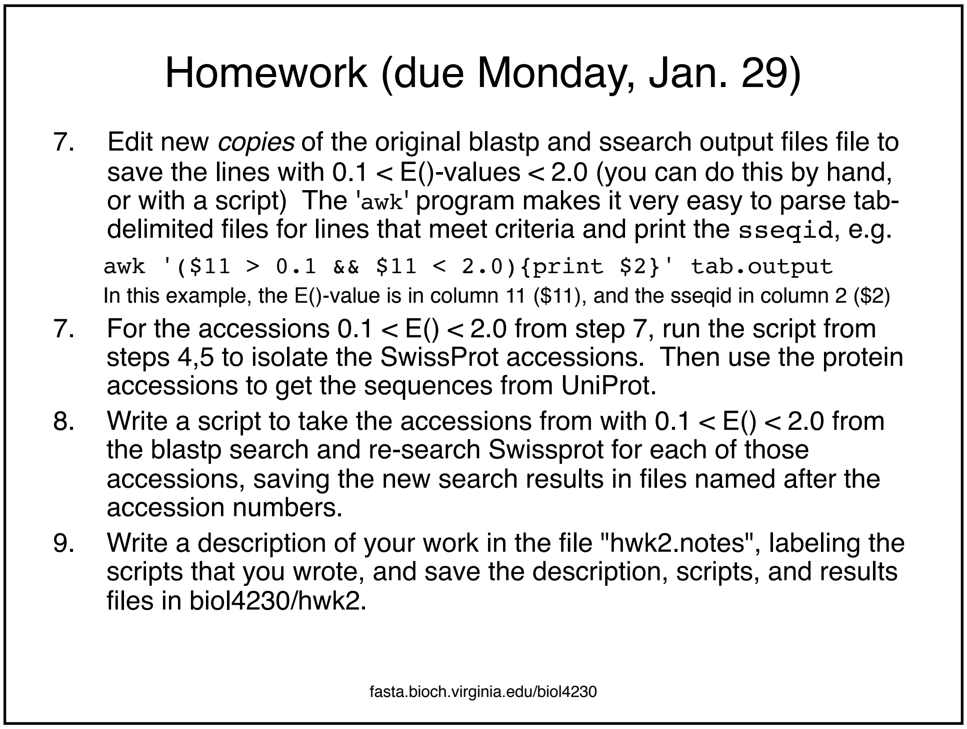


## Homework (due Monday, Jan. 29, 12:00 noon)

1. Do a search of the SwissProt database using `blastp` using `NP_001171499 (honeybee_gst.aa)` saving the output in "tabular" format (-outfmt 6)
2. Repeat step 1, using the `ssearch36` program specifying the BLOSUM62 matrix (`-s BP62`). (you can produce tabular output using the `-m 8` option):

   ```
   ssearch36 -m 8 -s BP62 honeybee_gst.aa q > output
   ```
3. For both the blastp and ssearch results, make a *copy* of each results file and remove all the lines with E() > 0.001. Write a bash script to isolate the library (subject) accession information for each of the lines in the edited file, and save the accession in a new file
4. For each accession, split it into its component parts (hit 'man cut' to see how to change the delimiter).
   - Write a script to save the accessions (P12345.3) to a file, and isolate only the accessions without the version information.
5. Compare the list of SwissProt accessions with E() < 0.001 from BLASTP and SSEARCH. Which program finds more homologs? For the program that finds fewer homologs, what are the E()-values of those hits in the list of hits from the other program?

fasta.bioch.virginia.edu/biol4230



## Homework (due Monday, Jan. 29)

7. Edit new *copies* of the original blastp and ssearch output files file to save the lines with 0.1 < E()-values < 2.0 (you can do this by hand, or with a script) The '`awk`' program makes it very easy to parse tab-delimited files for lines that meet criteria and print the `sseqid`, e.g.

   ```
   awk '($11 > 0.1 && $11 < 2.0){print $2}' tab.output
   ```
   In this example, the E()-value is in column 11 ($11), and the sseqid in column 2 ($2)

7. For the accessions 0.1 < E() < 2.0 from step 7, run the script from steps 4,5 to isolate the SwissProt accessions. Then use the protein accessions to get the sequences from UniProt.
8. Write a script to take the accessions from with 0.1 < E() < 2.0 from the blastp search and re-search Swissprot for each of those accessions, saving the new search results in files named after the accession numbers.
9. Write a description of your work in the file "hwk2.notes", labeling the scripts that you wrote, and save the description, scripts, and results files in biol4230/hwk2.

fasta.bioch.virginia.edu/biol4230