Introduction to command-line tools and programming software

S. *Ravi*chandran
ravichandran@hood.edu

Agenda

- Brief discussion about the Watson-Crick Paper
- Find-a-gene project (choice of gene)
- Biostars book
  - I Installation
  - II Unix
- R/RStudio; COLAB
- Edirect
  - Hood Cluster

http://dosequis.colorado.edu/Courses/MethodsLogic/papers/WatsonCrick1953.pdf

Discussion

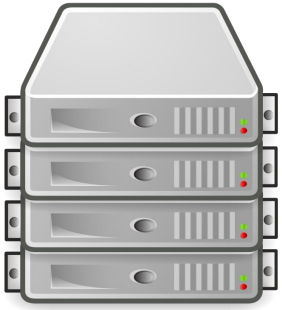
- What is the key difference between Linus Pauling's DNA model and Watson and Crick's model?

- What was the key conclusion of Watson and Crick's paper?

https://github.com/ravichas/bioinformatics

Windows users: Putty (installation required)

Mac: Terminal (no installation)



Linux Operating System
R (command line)
Other Bioinformatics software

R
RStudio
Jupyter → Python

Binder (Stopped working recently)

Hood Cluster

- User account?
  - Have you tried logging in?
- Need help, please contact Rob

Hood Cluster Login Details

We will be using the Hood Linux cluster called slurm1 for the hands-on exercises. If you do not have an account, please email Robert Jones, our system administrator, and request to create one.

Please note that if your edirect installation is working correctly, ignore these instructions.

- log into the cluster (Mac: ssh yourusername@slurm.hood.edu Windows: Use Putty to login)
- login into a node by typing the following line
  srun --pty bash -i
- run the command provided under the installation section of the following link, https://www.ncbi.nlm.nih.gov/books/NBK179288/
- The previous step will do the following two things:
  - Create **edirect** directory (with edirect scripts) under your home directory
  - Add couple of lines to your .bashrc file
- Make sure you are in the home directory. You can go to HOME directory by typing "cd"
- Test the installation with the following command:
  - edirect -db pubmed -query "asthma"

Unix/Linux

- **Warning**
- This is just a basic introduction to the Linux OS. For newcomers, it needs several weeks of exposure to become comfortable with this OS
- Today, we are just going to barely scratch the surface

Basic Intro to Shell

- Shell is a program that sits between the User and the OS.
- There are many flavors
  - C, K, Bourne, Bourne-again etc.
- C or tcsh is very user-friendly but rarely used by system administrators.
- I mostly use C-shell (tcsh)
- **I will demonstrate the capability in the class**

Entrez Direct (Edirect)

- Advanced method for accessing (not readily available via web-interface) the NCBI's databases
- Perl scripts accessible via Linux/Unix Shell scripts
- Will run on Unix/Linux/Mac systems that has Perl installed
  - If you are using Windows; use Cygwin (not complicated and I will not cover this in the class)

What is needed to run Edirect?

- Unix/Linux/Mac OSX
  - Windows
    - Unix/Linux emulators (SeqWin etc.)
- Perl with LWP::Simple
- Basic Idea
  - Each script outputs an XML
    - These XML outputs can in turn be piped using other Edirect commands
      - ESearch, Efetch, Elink, Epost xtract etc.
        - » Xtract (Powerful XML parser)

Why XML?
Easy to parse an XML
NCBI adopted XML based output

Where do we get this software?

NCBI Resources How To

NCBI
National Center for Biotechnology Information

All Databases Search

NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News | Blog

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

Popular Resources
PubMed
Bookshelf
PubMed Central
PubMed Health
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

NCBI Announcements
August 3rd webinar: NCBI T
Loci: RefSeq Ribosomal RN
Sequences for Identificatio
Phylogenetic Analysis

Tree Viewer 1.10 visualizes
phylogenetic trees up to 10

Tree Viewer version 1.10 ha

FEEDBACK

Develop

NCBI provides a variety of resources that allow developers to access and manipulate NCBI data in their applications.

APIs

Programming interfaces including the E-utilities, BLAST URL API and PUG

Code Libraries

Software libraries including the C++ and SRA toolkits

Data Formats

Schema, DTDs and other data specifications

GitHub

NCBI GitHub repository

News & Blog

Entrez Utilities News
C++ toolkit News
API blog posts

tps://www.ncbi.nlm.nih.gov/home/develop/api/

http://www.google.cc | IRF5 - Wikipedia, the | ScienceProject | Web Services provide | Free Online Course M | Google | Inbox - Web Email Ac | Cbl and human myel | Add Site to

NIH U.S. National Library of Medicine | NCBI National Center for Biotechnology Information | Sign in to NCBI

NCBI HOME LITERATURE HEALTH GENOMES GENES PROTEINS CHEMICALS POPULAR RESOURCES ▼

All Databases ▾ Search NCBI Search

APIs

NCBI provides several public APIs that allow programmatic access to many databases and tools.

Entrez Programming Utilities (E-utilities)

The E-utilities are the public API to the NCBI Entrez system and allow access to all Entrez databases including PubMed, PMC, Gene, Nuccore and Protein. The E-utilities are a suite of eight server-side programs that accept a fixed URL syntax for search, link and retrieval operations. A companion package named Entrez Direct consists of several executables that allow the E-utilities to be called directly from a UNIX command line.

Documentation Quick Start Examples Entrez Direct

BLAST URL API

The BLAST API allows developers to submit BLAST searches for processing at NCBI or cloud service provider(s) using HTTPS. The API can then check the status of submitted searches and retrieve results when ready in several formats.

Overview Documentation

PubChem Power User Gateway (PUG)

PUG is a suite of APIs for the NCBI PubChem resource, and provides programmatic access to many PubChem functions including downloads of chemical and assay data, chemical structure searches and chemical standardization. PUG offers HTTP, REST and SOAP interfaces, and also integrates with the E-utilities to provide convenient access to other NCBI databases.

Documentation SOAP REST REST Tutorial

PubMed Central (PMC) APIs

PMC provides several APIs that provide programmatic access to various services that deal with PMC literature content, including file validation tools, Open Access web services, and an ID convertor that interconverts PMCID's, PMID's, Manuscript ID's, and DOI's.

Documentation

ADDITIONAL LINKS

Developer guidance about HTTPS at NCBI

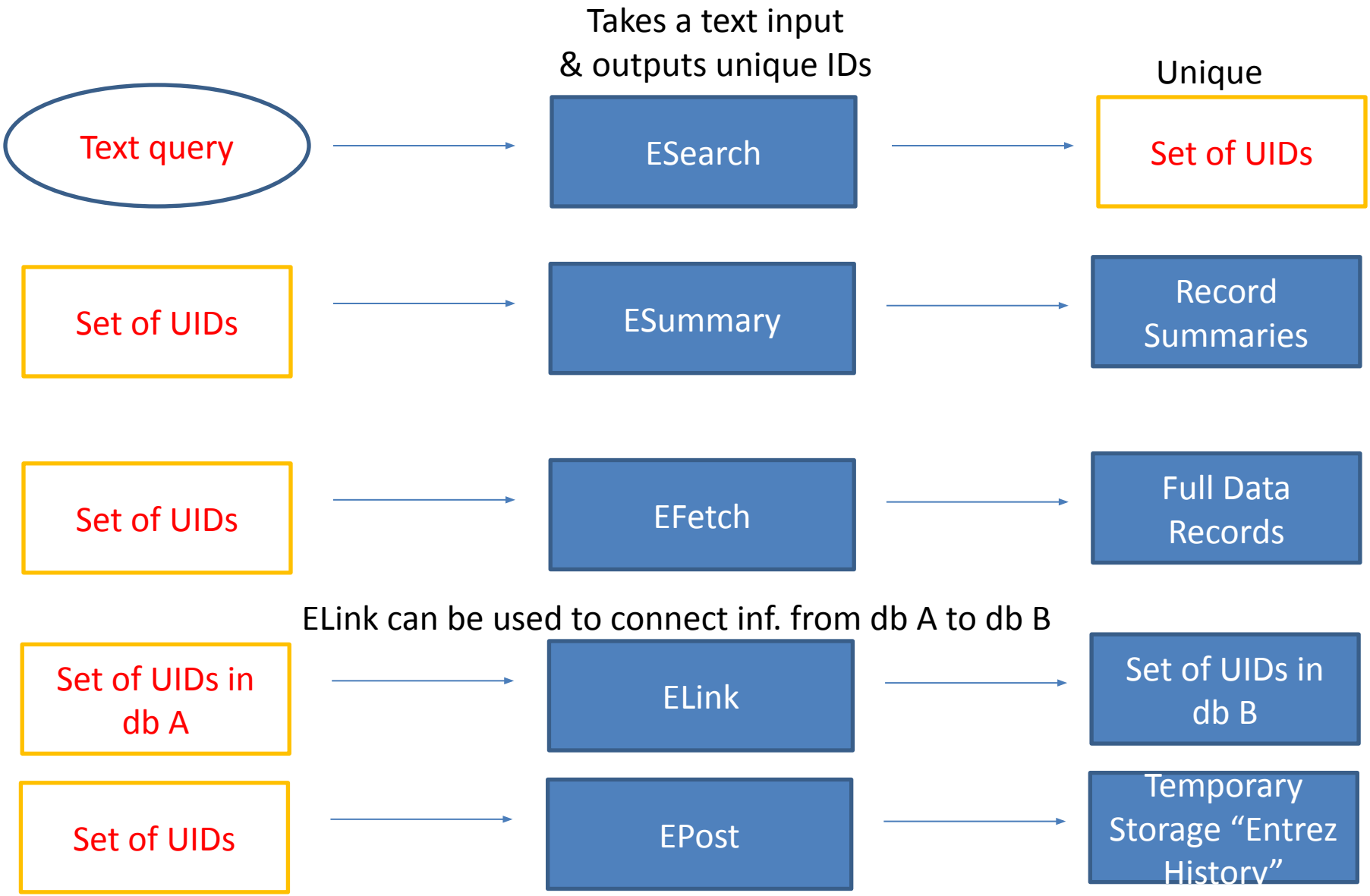
API Usage Guidelines

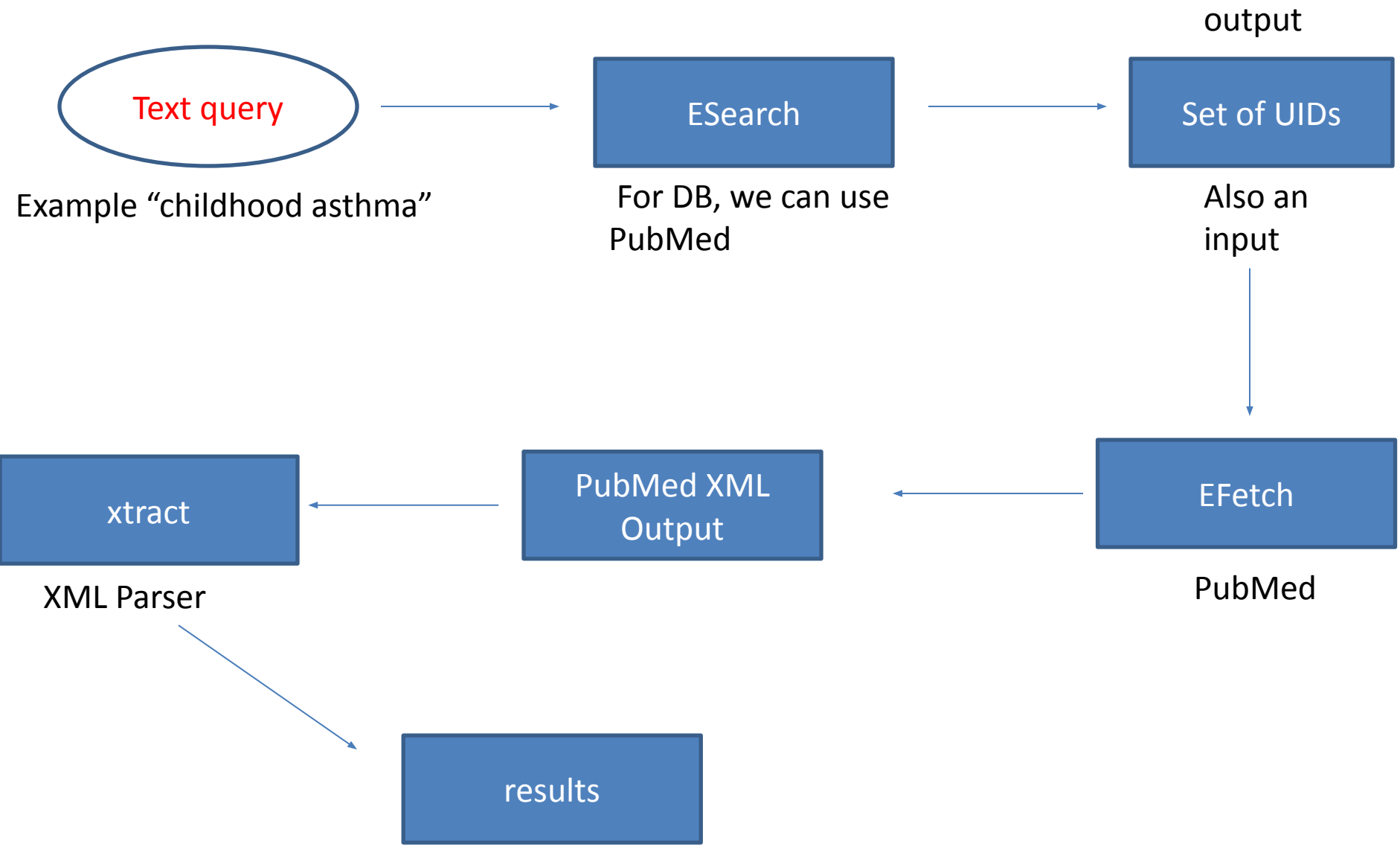
Data Usage Policy

NLM APIs

Based on NCBI Help Pages

E-utilities




output
Text query
ESearch
Set of UIDs
Example “childhood asthma”
For DB, we can use PubMed
Also an input
xtract
PubMed XML Output
EFetch
XML Parser
PubMed
results

Edirect/E-utilies

- Why is it useful?
- Let us do a couple of examples

NOTE: DO NOT USE QUOTES FOR THE STRING THAT YOU WILL USE FOR THE SEARCH

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed ▾ childhood asthma

Create RSS Create alert Advanced

Article types
Clinical Trial
Review
Customize ...

Text availability
Abstract
Free full text
Full text

PubMed Commons
Reader comments
Trending articles

Publication dates
5 years
10 years
Custom range...

Format: Summary ▾ **Sort by**: Most Recent ▾ **Send to** ▾

 # WILL BE DIFFERENT

Search results

Items: 1 to 20 of 10039

<< First < Prev Page 1 of 502 Next > Last >>

1. Early Life Exposure to the Great Smog of 1952 and the Development of **Asthma**.
Bharadwaj P, Graff Zivin J, Mullins JT, Neidell M.
Am J Respir Crit Care Med. 2016 Jul 8. [Epub ahead of print]
PMID: 27392261

2. Domestic Dog Exposure at birth reduces the Incidence of Atopic Dermatitis.
Thorsteinsdottir S, Thyssen JP, Stokholm J, Vissing NH, Waage J, Bisgaard H.
Allergy. 2016 Jul 6. doi: 10.1111/all.12980. [Epub ahead of print]

Let us see how we can use E-Direct to get the same output using Command Prompt

esearch -db pubmed -query "childhood asthma"

```
<ENTREZ_DIRECT>
  <Db>pubmed</Db>
  <WebEnv>NCID_1_90103729_130.14.18.34_9001_1468093986_406367782_0MetA0_S_MegaStore_F_1</WebEnv>
  <QueryKey>1</QueryKey>
  <Count>10039</Count>
  <Step>1</Step>
</ENTREZ_DIRECT>
```

Look at the count, it is same as the web based search.

Is it useful? It just gives you a total count. If you want individual records, you have to use esummary

esearch -db pubmed -query "childhood asthma“ | esummary

Note this will keep dumping all the 10,039 records in the XML format.

Edirect/E-utilies

- Second example
- For example, if you want to search PubMed for articles that contain the keyword, “lycopene cyclase” and download all the resulting article abstracts into a file.

Applications of E-Direct

- Go to PubMed, type “lycopene cyclase”
  - Hit “**Search**”
- From the hits page, change the Format of the page to **“Abstract”** and also ask it to show 200 hits per page to see them all
- Use **Send to** a file for download
- **E-Utils**
- **esearch -db pubmed -query "lycopene cyclase" | \ efetch -format abstract > HitsFile.txt**

- esearch –db protein –query “lycopene cyclase” | efetch –format fasta
- Please read more on the following NCBI link:
- http://www.ncbi.nlm.nih.gov/books/NBK179288/

Hints:

- use
  - esearch -help (for options)

Key flags that work well with xtract

- **-sep** and **–tab**

- **esearch –db pubmed –query "LLM"**
  - **Will get a XML output**
  - **Which is a summary file # of hits etc,**
  - **Not very useful; to extract summaries, we pipe it to esummary**
- **esearch –db pubmed –query "LLM" | esummary**

```
esearch –db pubmed –query “LLM” \
  | efetch –format xml | xtract –pattern  \
PubmedArticle  -block Author –element \
LastName, Initials
```

-pattern Pubmed Article will loop over each article

Within each article, block will search

Output will be a bunch of names LastName, Initials separated by tab character (default).

To force the script to use space, use “-sep” flag

```
esearch –db pubmed –query “LLM” \
  | efetch –format xml | xtract –pattern  \ PubmedArticle  -block \
Author –sep “ “ –element \ LastName, Initials
```

If we want each Author in a new line

```
esearch –db pubmed –query “LLM” \
  | efetch –format xml | xtract –pattern  \ PubmedArticle  -block \
Author –sep “ “ –tab “\n” –element \ LastName, Initials
```

? How do we store the results in a file? (Linux shell command)

Take that file and sort them and see the frequency of each author

Hint we will use two shell command: sort and uniq

Links

- Linux
  - https://web.stanford.edu/group/farmshare/cgi-bin/wiki/index.php/How_to_learn_linux
- NCBI E-utilities
  - API to NCBI Entrez system
  - Need: Internet, web-browser, Perl etc.
  - http://www.ncbi.nlm.nih.gov/books/NBK25501/
- R
  - https://www.r-project.org/ (Windows/Mac/Linux)
- Rstudio
  - https://www.rstudio.com/ (Windows/Mac/Linux)
- https://github.com/ravichas/bioinformatics

Statistics

- R/Rstudio
- R
  - Free Software for statistical computing and graphics.
  - It compiles and runs on a wide variety of OS
    - Unix, Windows, MacOS. UNIX platforms, Windows and MacOS.
- RStudio
  - GUI for R; Very helpful to run R via RStudio

Edirect

```
esearch -db nucleotide -query PRJNA257197 | efetch -format fasta > genomes.fa
```

```
You can run in your local laptop as well as in Hood Cluster
```