Introduction to Bioinformatics

Alejandro Rojas

PLPA504V/CEMB590V

About today

- Get to known each other

- Course organization

- Course outline and goals

Course goals

- After this class you should be able to:
  - Use some basic bioinformatics tools and choose the right parameters.
  - Handle large-scale datasets with computer programming.
  - Resolve biological problems with bioinformatics.

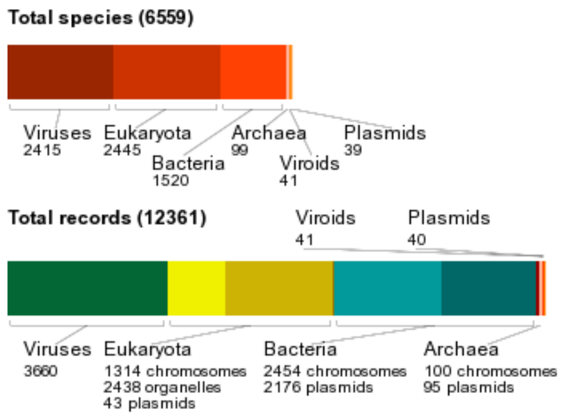
Two important questions

- Why is it a good thing that you are learning bioinformatics?
- What is bioinformatics anyway?
  - To answer these questions, we need to first talk about biological data.

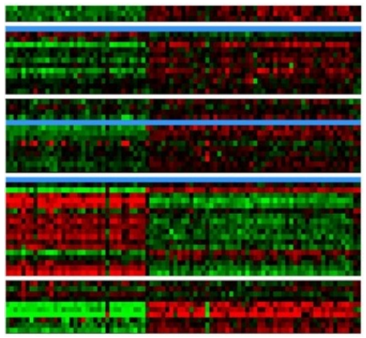
- What kinds of biological data are available?

Avalanche of biological data

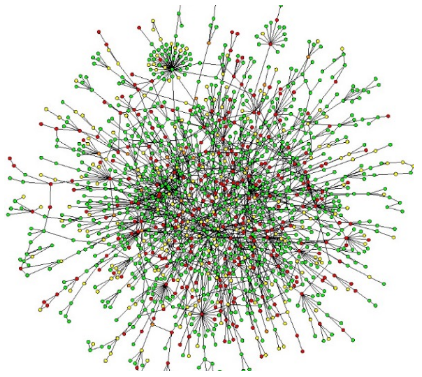
- You need to know how to play with these data with *computation and bioinformatics*.



>6000 genome projects with some records in NCBI.



469,860 expression experiments in NCBI GEO



At least one interactions for 2.5 million proteins from 630 organisms in the String database

Bioinformatics

- Research, development, or application of
- Computational tools and approaches for
- Expanding the use of biological, medical, behavioral, health data,
- Including those to acquire, store, organize, archive, analyze, or visualize such data.

http://www.bisti.nih.gov/CompuBioDef.pdf

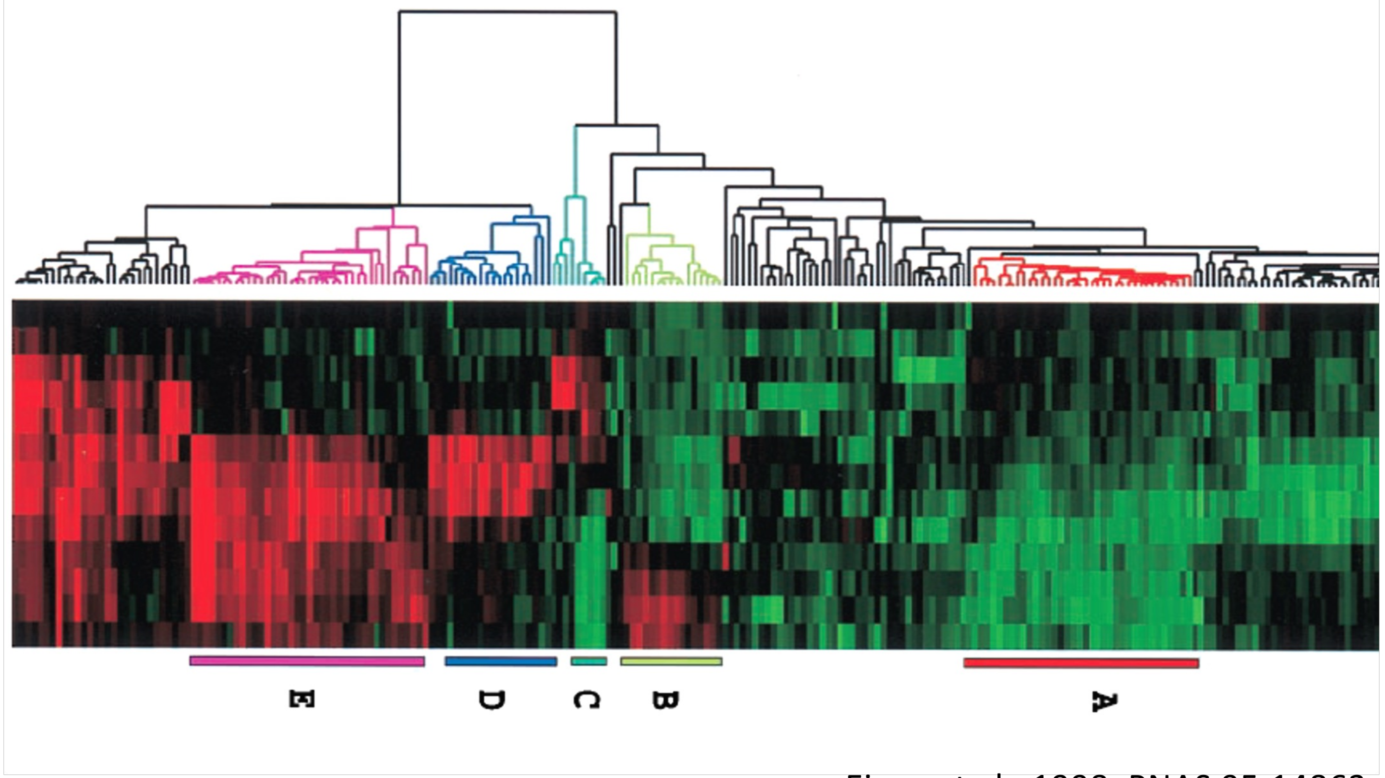
Algorithm, Data Structure, Programming

- Two independent stress-treatments
  - Lists of differentially expressed genes. How will you find out the common genes?

List 1	***List 2***
Thx	Lnn1
Phh1	Thx
Lok4	Thy
Jij3	Phh1
Arg25	Kwe40
Arg2	Arg2
Qrt3	Arg3
Kww4	Jij3
	Arg25

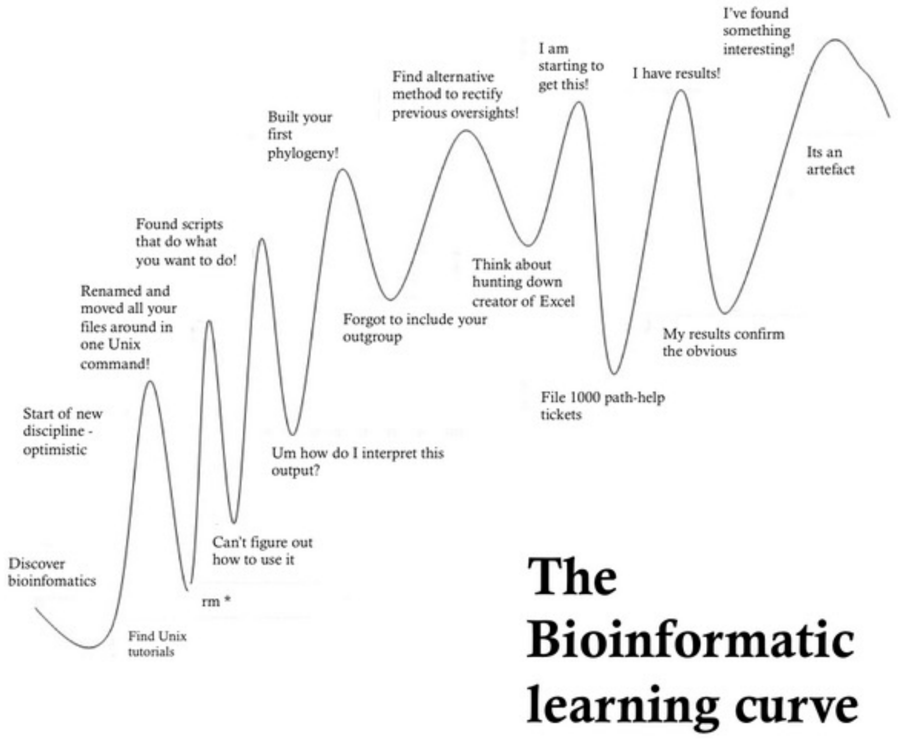
Artificial Intelligence & Robotics

- How can we identify the underlying structure in biological data?

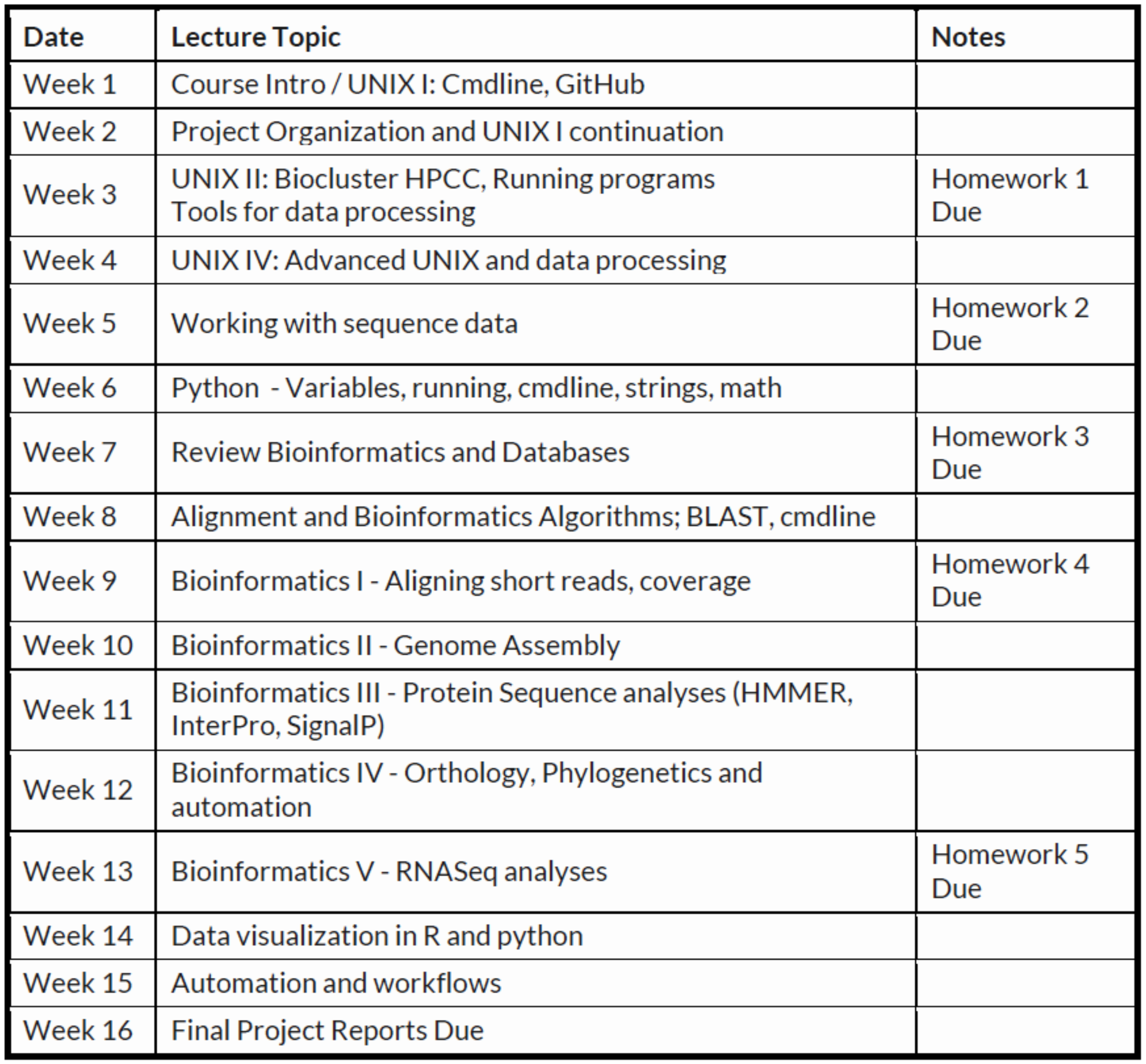


Eisen et al., 1998. PNAS 95:14863

Steep learning curve

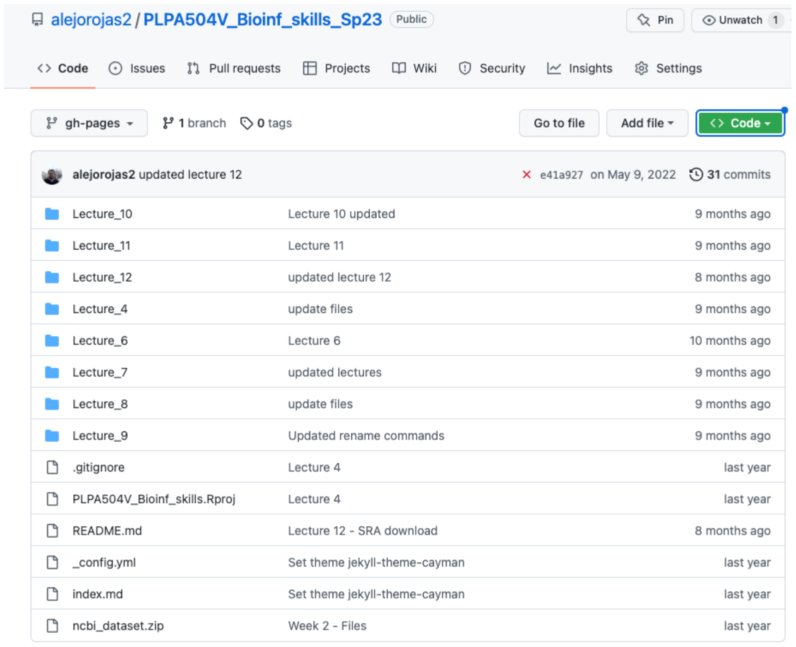

Discover bioinfomatics
Find Unix tutorials
Start of new discipline - optimistic
Renamed and moved all your files around in one Unix command!
rm *
Found scripts that do what you want to do!
Can't figure out how to use it
Built your first phylogeny!
Um how do I interpret this output?
Forgot to include your outgroup
Find alternative method to rectify previous oversights!
Think about hunting down creator of Excel
I am starting to get this!
File 1000 path-help tickets
I have results!
My results confirm the obvious
I've found something interesting!
Its an artefact
The Bioinformatic learning curve

Class schedule



Date	Lecture Topic	Notes
Week 1	Course Intro / UNIX I: Cmdline, GitHub	
Week 2	Project Organization and UNIX I continuation	
Week 3	UNIX II: Biocluster HPCC, Running programs Tools for data processing	Homework 1 Due
Week 4	UNIX IV: Advanced UNIX and data processing	
Week 5	Working with sequence data	Homework 2 Due
Week 6	Python - Variables, running, cmdline, strings, math	
Week 7	Review Bioinformatics and Databases	Homework 3 Due
Week 8	Alignment and Bioinformatics Algorithms; BLAST, cmdline	
Week 9	Bioinformatics I - Aligning short reads, coverage	Homework 4 Due
Week 10	Bioinformatics II - Genome Assembly	
Week 11	Bioinformatics III - Protein Sequence analyses (HMMER, InterPro, SignalP)	
Week 12	Bioinformatics IV - Orthology, Phylogenetics and automation	
Week 13	Bioinformatics V - RNASeq analyses	Homework 5 Due
Week 14	Data visualization in R and python	
Week 15	Automation and workflows	
Week 16	Final Project Reports Due	

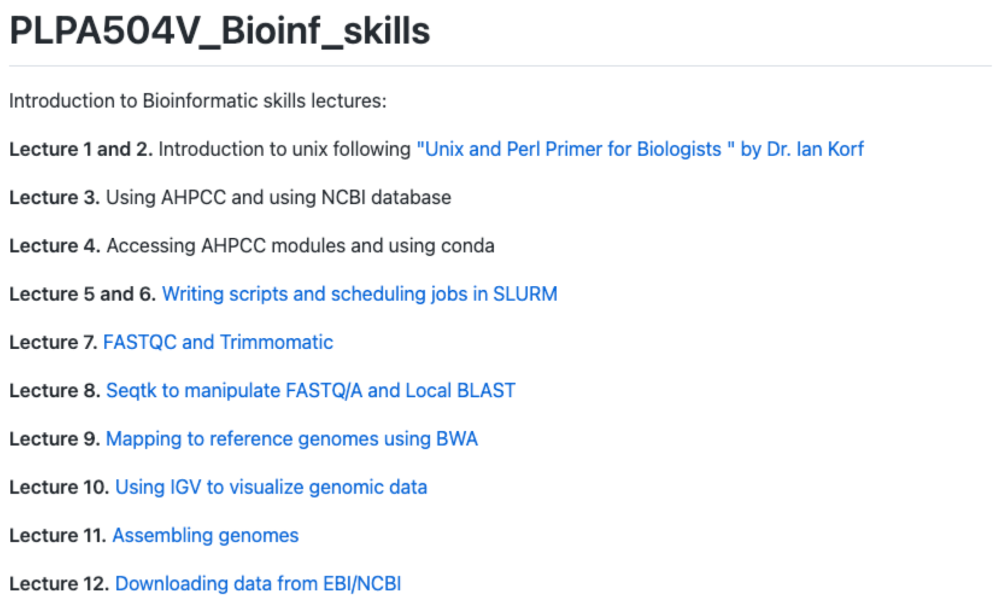
Class material



alejorojas2 / PLPA504V_Bioinf_skills_Sp23 Public

alejorojas2 updated lecture 12 — e41a927 on May 9, 2022 — 31 commits

Lecture_10	Lecture 10 updated	9 months ago
Lecture_11	Lecture 11	9 months ago
Lecture_12	updated lecture 12	8 months ago
Lecture_4	update files	9 months ago
Lecture_6	Lecture 6	10 months ago
Lecture_7	updated lectures	9 months ago
Lecture_8	update files	9 months ago
Lecture_9	Updated rename commands	9 months ago
.gitignore	Lecture 4	last year
PLPA504V_Bioinf_skills.Rproj	Lecture 4	last year
README.md	Lecture 12 - SRA download	8 months ago
_config.yml	Set theme jekyll-theme-cayman	last year
index.md	Set theme jekyll-theme-cayman	last year
ncbi_dataset.zip	Week 2 - Files	last year



PLPA504V_Bioinf_skills

Introduction to Bioinformatic skills lectures:

Lecture 1 and 2. Introduction to unix following "Unix and Perl Primer for Biologists " by Dr. Ian Korf

Lecture 3. Using AHPCC and using NCBI database

Lecture 4. Accessing AHPCC modules and using conda

Lecture 5 and 6. Writing scripts and scheduling jobs in SLURM

Lecture 7. FASTQC and Trimmomatic

Lecture 8. Seqtk to manipulate FASTQ/A and Local BLAST

Lecture 9. Mapping to reference genomes using BWA

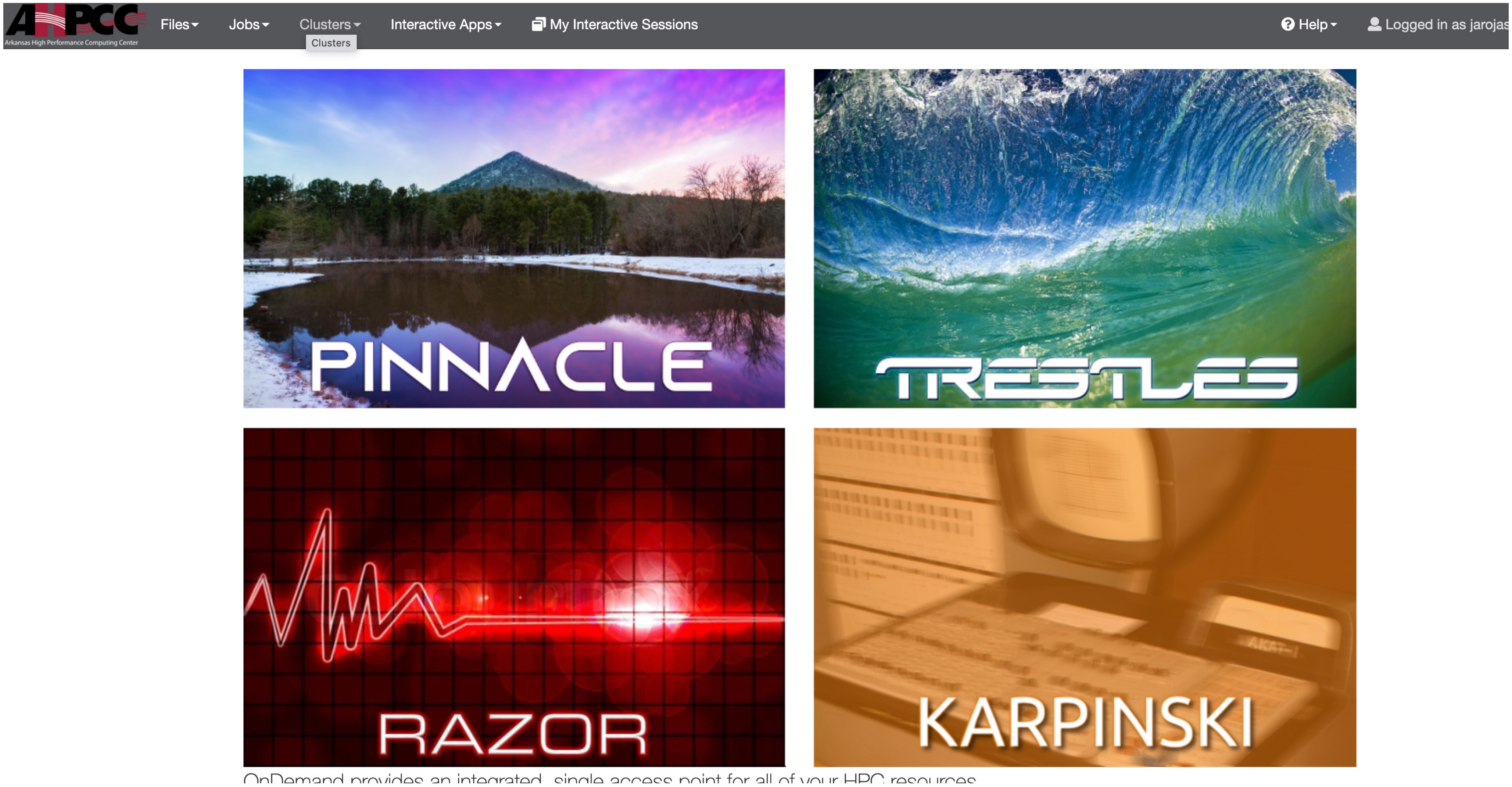
Lecture 10. Using IGV to visualize genomic data

Lecture 11. Assembling genomes

Lecture 12. Downloading data from EBI/NCBI

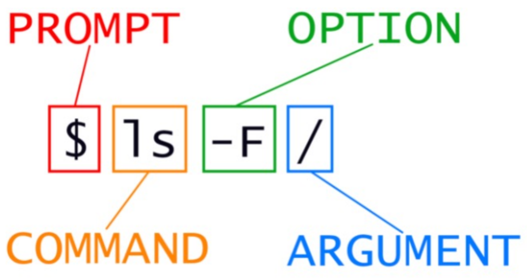
https://github.com/alejorojas2/PLPA504V_Bioinf_skills_Sp23

Review Shell and Unix



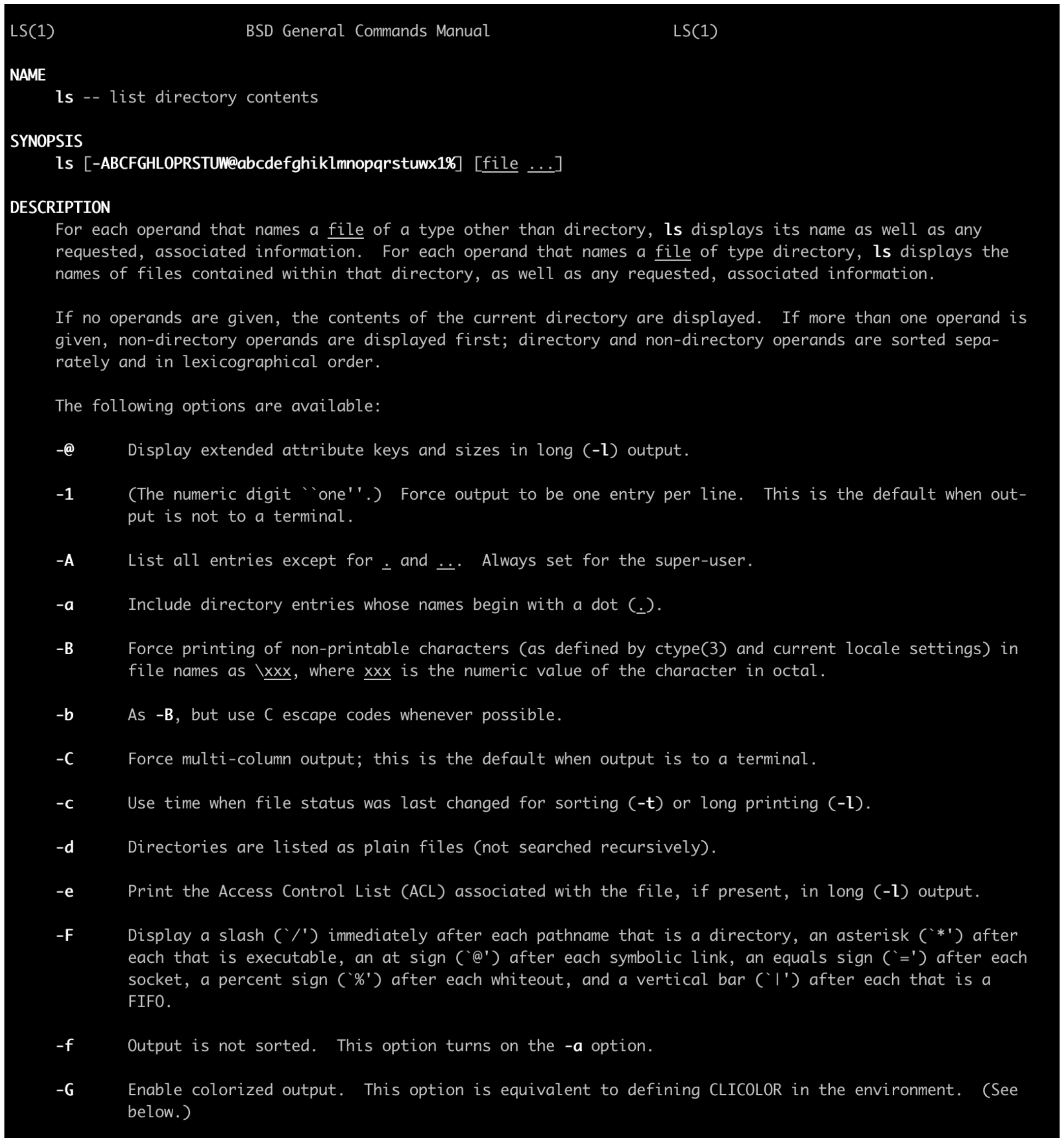
OnDemand provides an integrated, single access point for all of your HPC resources.

https://pinnacle-portal.uark.edu/

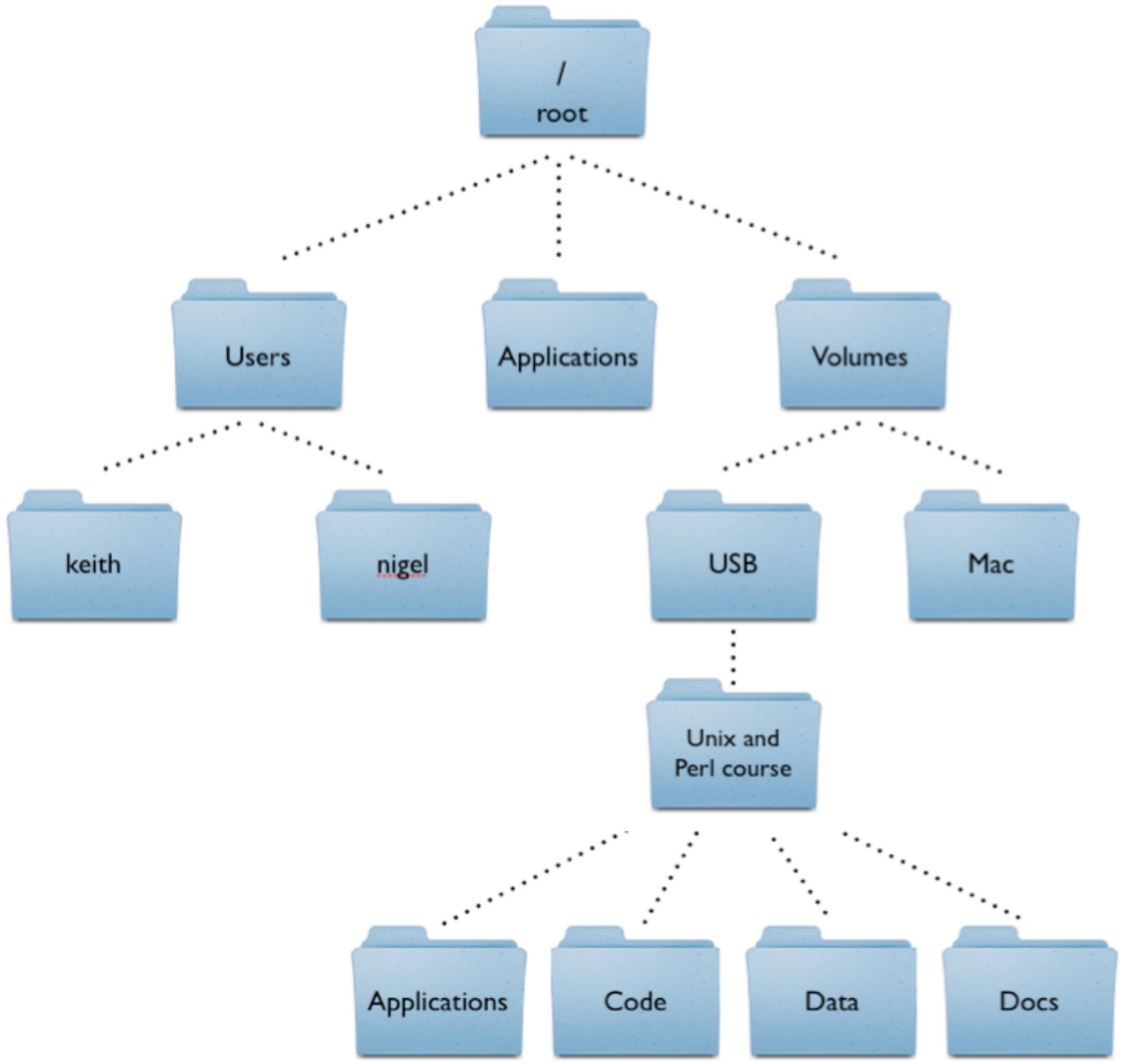


Getting help!

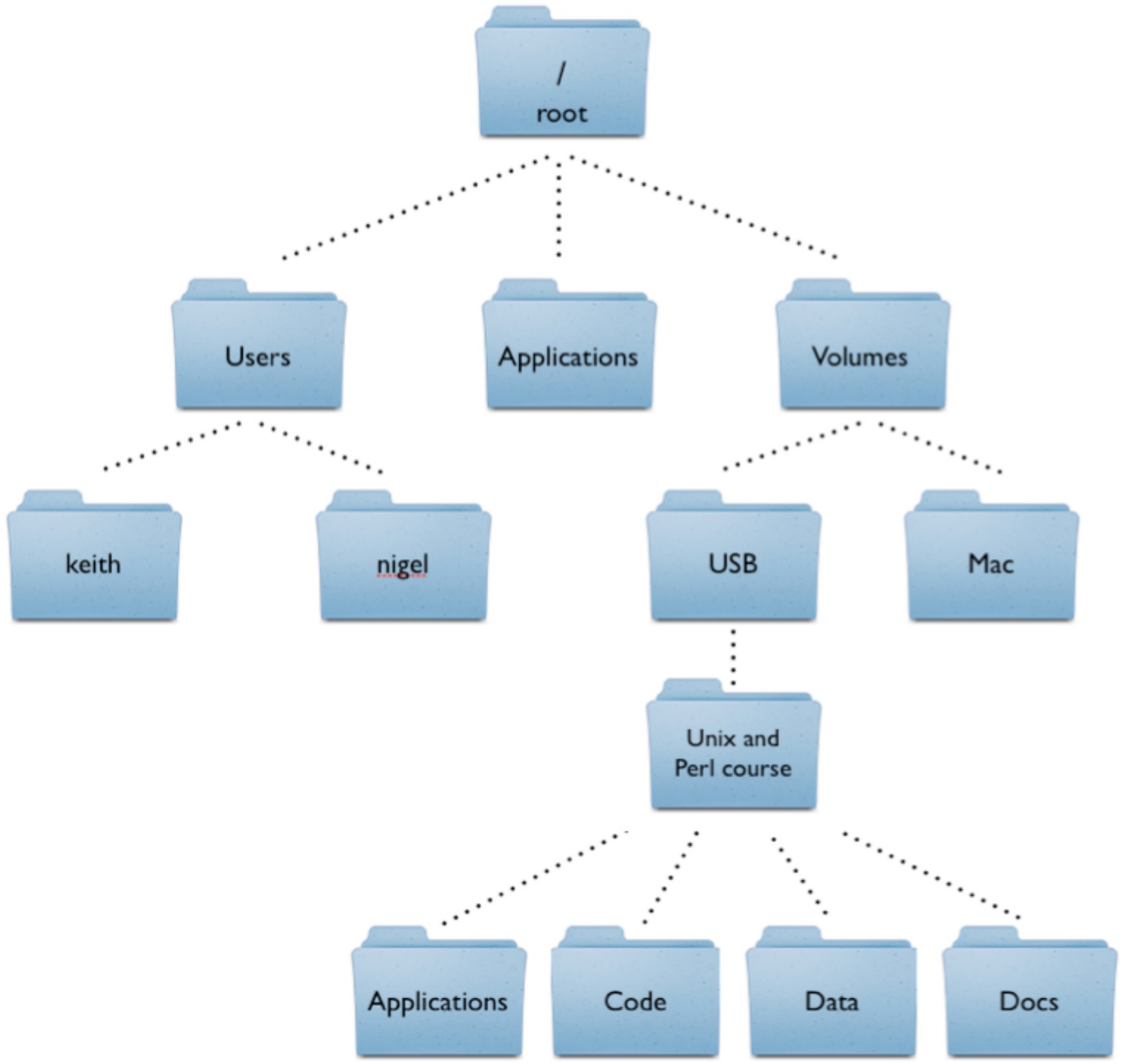
- $ man ls
- Understand syntax of the command



```
LS(1)                   BSD General Commands Manual                  LS(1)

NAME
     ls -- list directory contents

SYNOPSIS
     ls [-ABCFGHLOPRSTUW@abcdefghiklmnopqrstuwx1%] [file ...]

DESCRIPTION
     For each operand that names a file of a type other than directory, ls displays its name as well as any
     requested, associated information.  For each operand that names a file of type directory, ls displays the
     names of files contained within that directory, as well as any requested, associated information.

     If no operands are given, the contents of the current directory are displayed.  If more than one operand is
     given, non-directory operands are displayed first; directory and non-directory operands are sorted sepa-
     rately and in lexicographical order.

     The following options are available:

     -@      Display extended attribute keys and sizes in long (-l) output.

     -1      (The numeric digit ``one''.)  Force output to be one entry per line.  This is the default when out-
             put is not to a terminal.

     -A      List all entries except for . and ...  Always set for the super-user.

     -a      Include directory entries whose names begin with a dot (.).

     -B      Force printing of non-printable characters (as defined by ctype(3) and current locale settings) in
             file names as \xxx, where xxx is the numeric value of the character in octal.

     -b      As -B, but use C escape codes whenever possible.

     -C      Force multi-column output; this is the default when output is to a terminal.

     -c      Use time when file status was last changed for sorting (-t) or long printing (-l).

     -d      Directories are listed as plain files (not searched recursively).

     -e      Print the Access Control List (ACL) associated with the file, if present, in long (-l) output.

     -F      Display a slash (`/') immediately after each pathname that is a directory, an asterisk (`*') after
             each that is executable, an at sign (`@') after each symbolic link, an equals sign (`=') after each
             socket, a percent sign (`%') after each whiteout, and a vertical bar (`|') after each that is a
             FIFO.

     -f      Output is not sorted.  This option turns on the -a option.

     -G      Enable colorized output.  This option is equivalent to defining CLICOLOR in the environment.  (See
             below.)
```



Example directory structure



Example directory structure

- **pwd -** present working directory
- **ls -** list files
  - ls -l (list files as a list)
  - ls -lht (list files in a list sorted by most recent and humans readable file size)
- **cd -** change directory
  - **cd .** - current directory
  - **cd ..** - previous directory
  - **cd ~** - go to home folder
- **mkdir** - make directory
- **mv –** move or rename files
- **cp –** copy files
- **rm –** remove or delete files

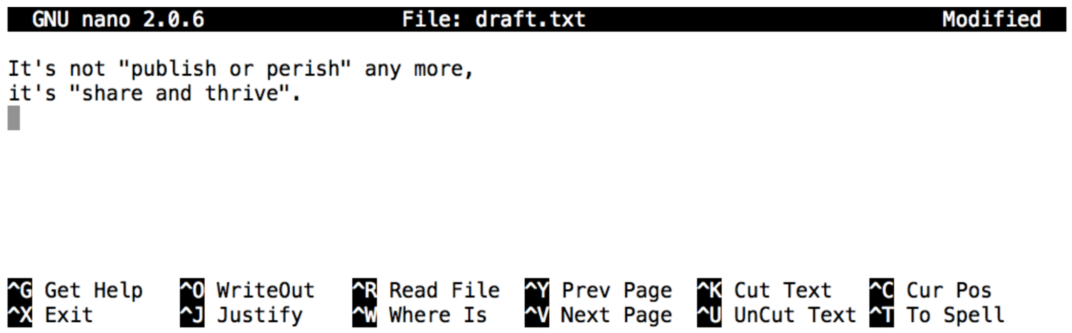
Text editor

- Nano, Vim, Emacs…
- $ nano draft.txt

- ^ = Ctrl

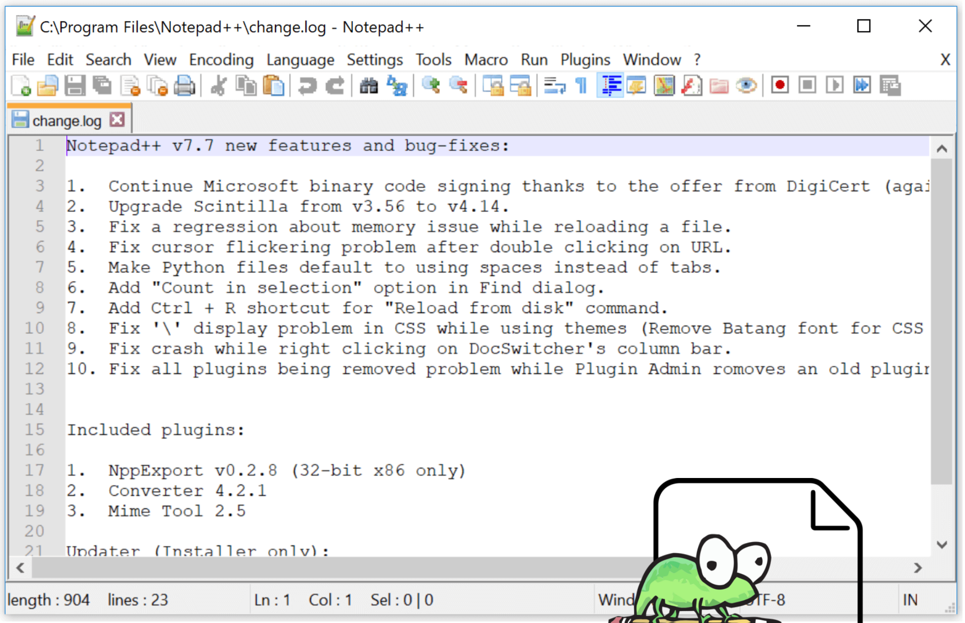
- Write or save

  Ctrl + O



```
  GNU nano 2.0.6                File: draft.txt                         Modified

It's not "publish or perish" any more,
it's "share and thrive".



^G Get Help  ^O WriteOut  ^R Read File ^Y Prev Page ^K Cut Text  ^C Cur Pos
^X Exit      ^J Justify   ^W Where Is  ^V Next Page ^U UnCut Text^T To Spell
```

Text editor

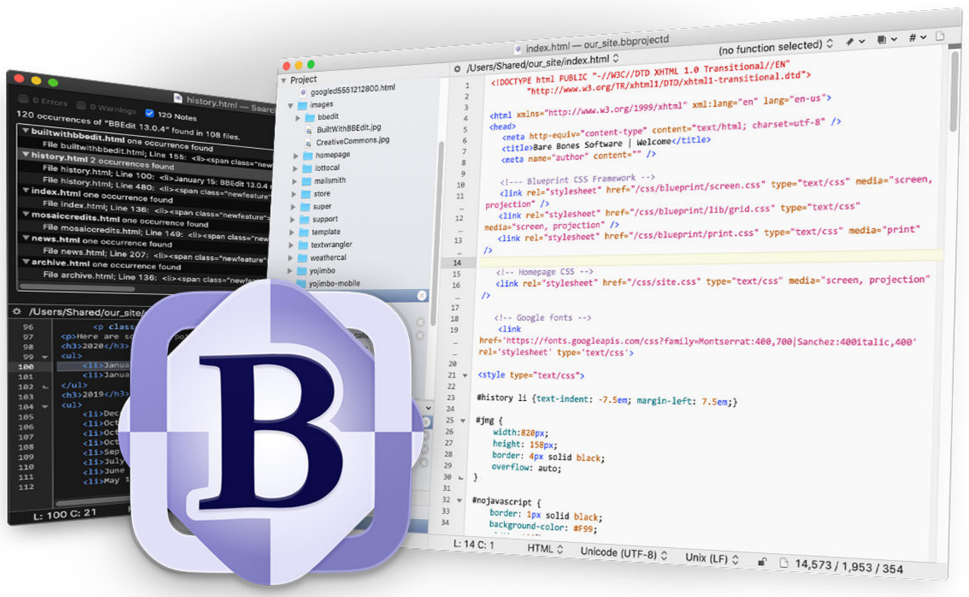


C:\Program Files\Notepad++\change.log - Notepad++

```
Notepad++ v7.7 new features and bug-fixes:

1.  Continue Microsoft binary code signing thanks to the offer from DigiCert (agai
2.  Upgrade Scintilla from v3.56 to v4.14.
3.  Fix a regression about memory issue while reloading a file.
4.  Fix cursor flickering problem after double clicking on URL.
5.  Make Python files default to using spaces instead of tabs.
6.  Add "Count in selection" option in Find dialog.
7.  Add Ctrl + R shortcut for "Reload from disk" command.
8.  Fix '\' display problem in CSS while using themes (Remove Batang font for CSS
9.  Fix crash while right clicking on DocSwitcher's column bar.
10. Fix all plugins being removed problem while Plugin Admin romoves an old plugir


Included plugins:

1.  NppExport v0.2.8 (32-bit x86 only)
2.  Converter 4.2.1
3.  Mime Tool 2.5

Updater (Installer only):
```

length : 904   lines : 23   Ln : 1   Col : 1   Sel : 0 | 0

Windows



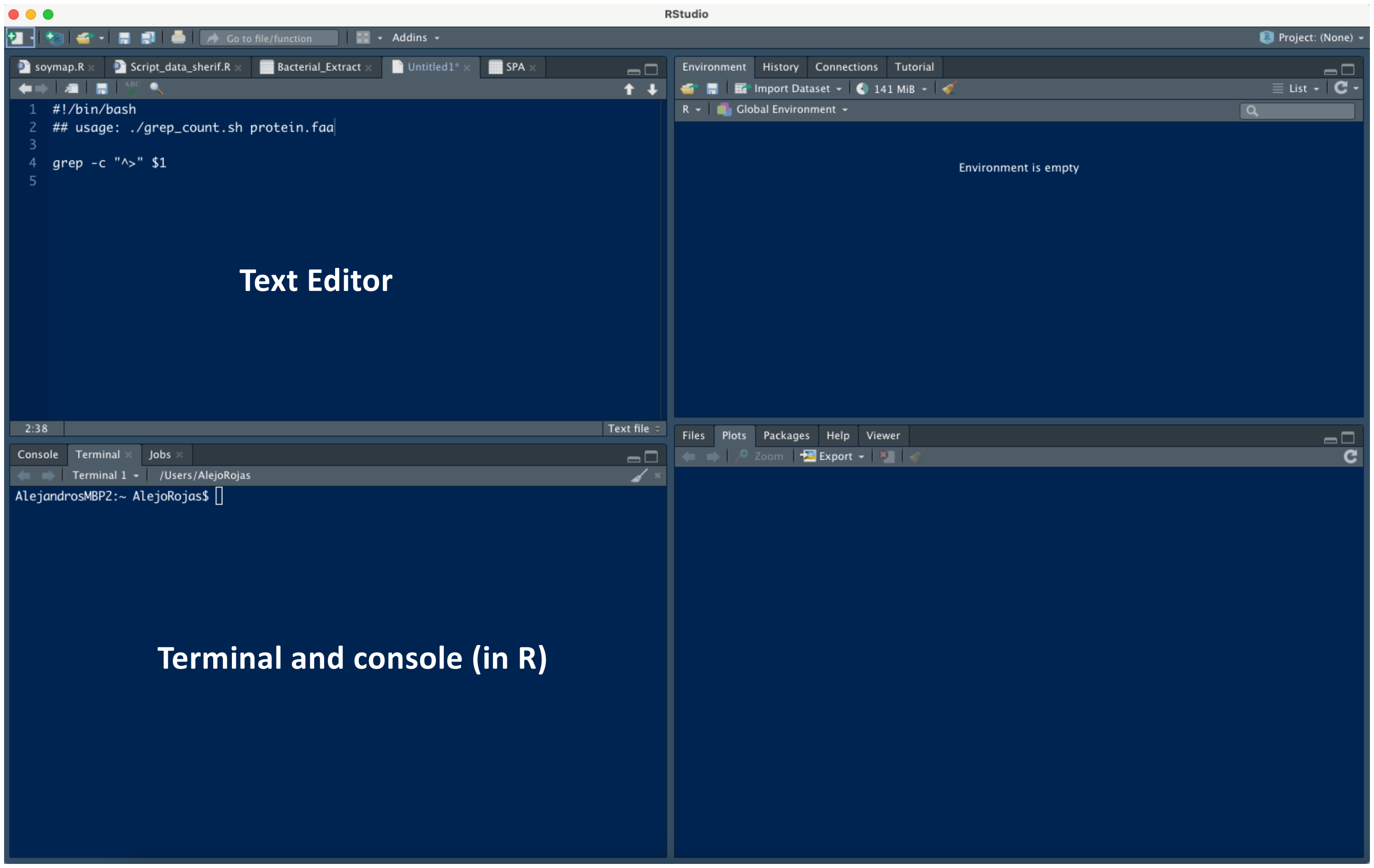
index.html — our_site.bbprojectd
/Users/Shared/our_site/index.html

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
        "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">

<html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en-us">
<head>
  <meta http-equiv="content-type" content="text/html; charset=utf-8" />
  <title>Bare Bones Software | Welcome</title>
  <meta name="author" content="" />

  <!--- Blueprint CSS Framework -->
  <link rel="stylesheet" href="/css/blueprint/screen.css" type="text/css" media="screen, projection" />
  <link rel="stylesheet" href="/css/blueprint/lib/grid.css" type="text/css" media="screen, projection" />
  <link rel="stylesheet" href="/css/blueprint/print.css" type="text/css" media="print" />

  <!-- Homepage CSS -->
  <link rel="stylesheet" href="/css/site.css" type="text/css" media="screen, projection" />

  <!-- Google fonts -->
  <link href='https://fonts.googleapis.com/css?family=Montserrat:400,700|Sanchez:400italic,400' rel='stylesheet' type='text/css'>

<style type="text/css">

#history li {text-indent: -7.5em; margin-left: 7.5em;}

#jmg {
   width:820px;
   height: 158px;
   border: 4px solid black;
   overflow: auto;
}

#nojavascript {
  border: 1px solid black;
  background-color: #F99;
```

L: 14 C: 1   HTML   Unicode (UTF-8)   Unix (LF)   14,573 / 1,953 / 354

Mac

IDEs (Interactive Development Environment)

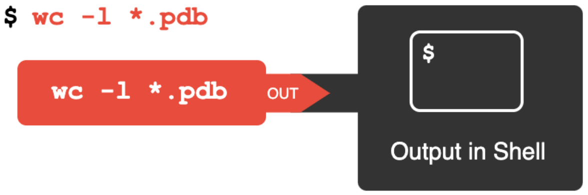


```
#!/bin/bash
## usage: ./grep_count.sh protein.faa

grep -c "^>" $1

```

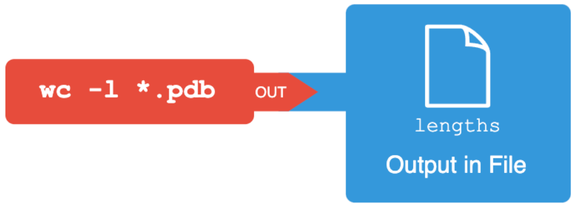
```
AlejandrosMBP2:~ AlejoRojas$
```

Merging commands and std output

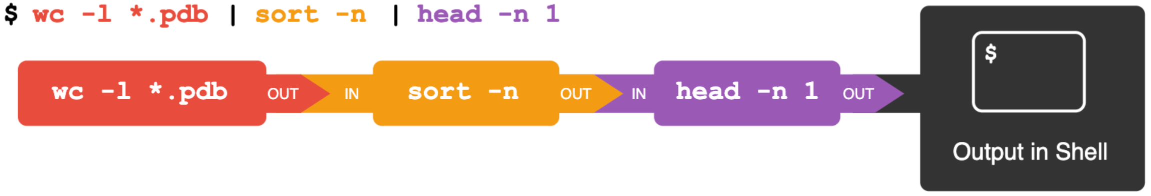
```
$ wc -l *.pdb
```



```
$ wc -l *.pdb > lengths
```



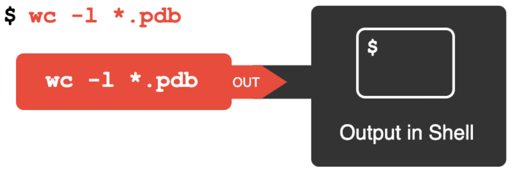
```
$ wc -l *.pdb | sort -n | head -n 1
```



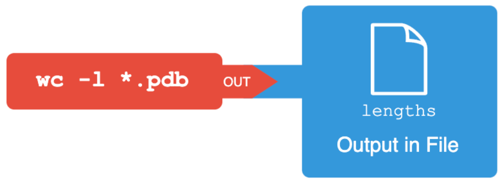
Merging commands and std output

- **wc -** counts lines, words, and characters in its inputs.
- **cat -** displays the contents of its inputs.
- **sort -** sorts its inputs.
- **head -** displays the first 10 lines of its input.
- **tail -** displays the last 10 lines of its input.
- **cut -** command is used to remove or ‘cut out’ certain sections of each line in the file
- **command > [file]** - redirects a command’s output to a file (overwriting any existing content).
- **command >> [file]** - appends a command’s output to a file.
- **[first] | [second]** is a pipeline: the output of the first command is used as the input to the second.

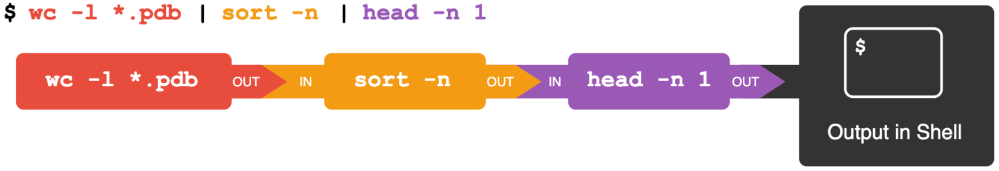
```
$ wc -l *.pdb
```

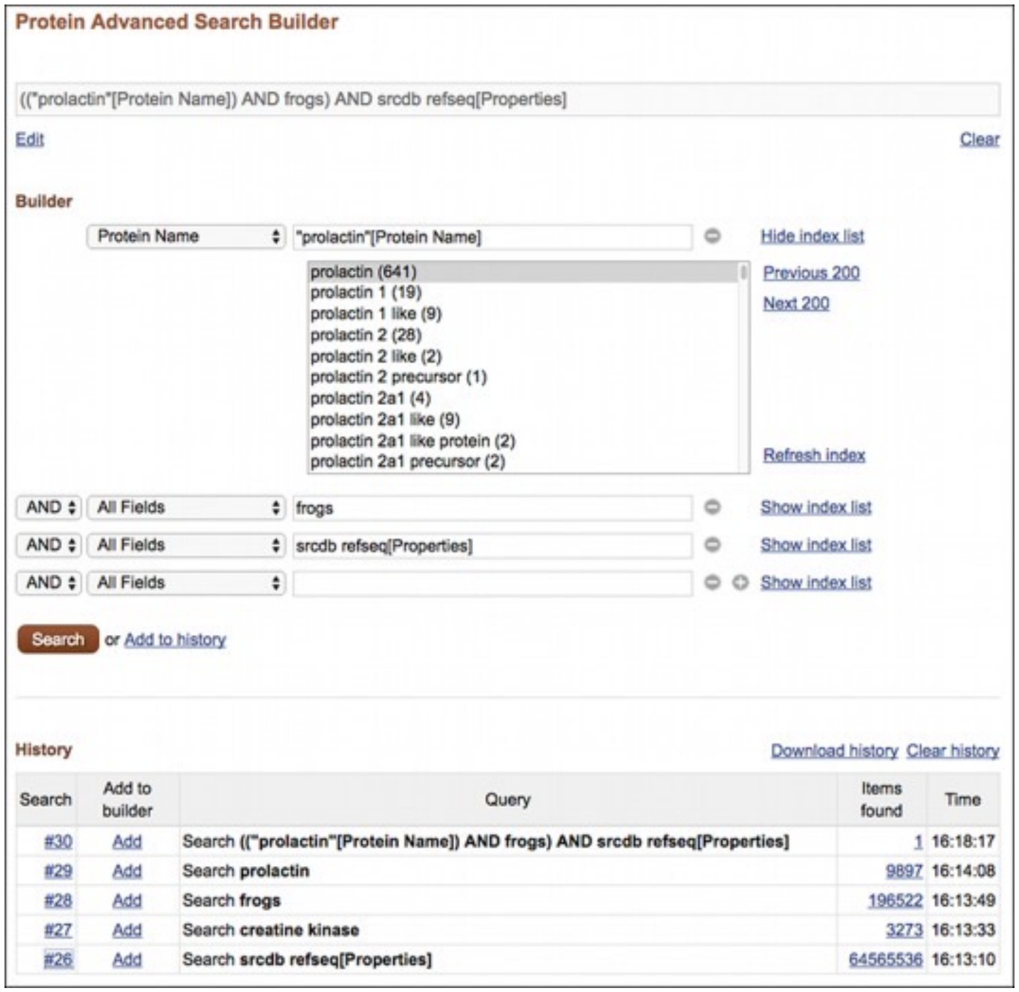


```
$ wc -l *.pdb > lengths
```



```
$ wc -l *.pdb | sort -n | head -n 1
```





Protein Advanced Search Builder

(("prolactin"[Protein Name]) AND frogs) AND srcdb refseq[Properties]

Edit Clear

Builder

	Protein Name	"prolactin"[Protein Name]	Hide index list
AND	All Fields	frogs	Show index list
AND	All Fields	srcdb refseq[Properties]	Show index list
AND	All Fields		Show index list

prolactin (641)
prolactin 1 (19)
prolactin 1 like (9)
prolactin 2 (28)
prolactin 2 like (2)
prolactin 2 precursor (1)
prolactin 2a1 (4)
prolactin 2a1 like (9)
prolactin 2a1 like protein (2)
prolactin 2a1 precursor (2)

Previous 200
Next 200
Refresh index

Search or Add to history

History

Download history Clear history

Search	Add to builder	Query	Items found	Time
#30	Add	Search (("prolactin"[Protein Name]) AND frogs) AND srcdb refseq[Properties]	1	16:18:17
#29	Add	Search prolactin	9897	16:14:08
#28	Add	Search frogs	196522	16:13:49
#27	Add	Search creatine kinase	3273	16:13:33
#26	Add	Search srcdb refseq[Properties]	64565536	16:13:10

https://www.ncbi.nlm.nih.gov/genbank/samplerecord/