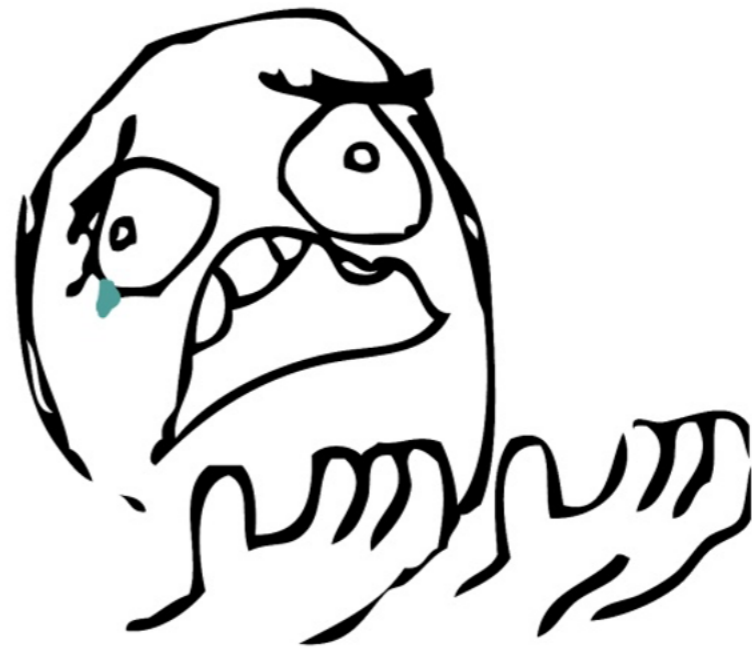


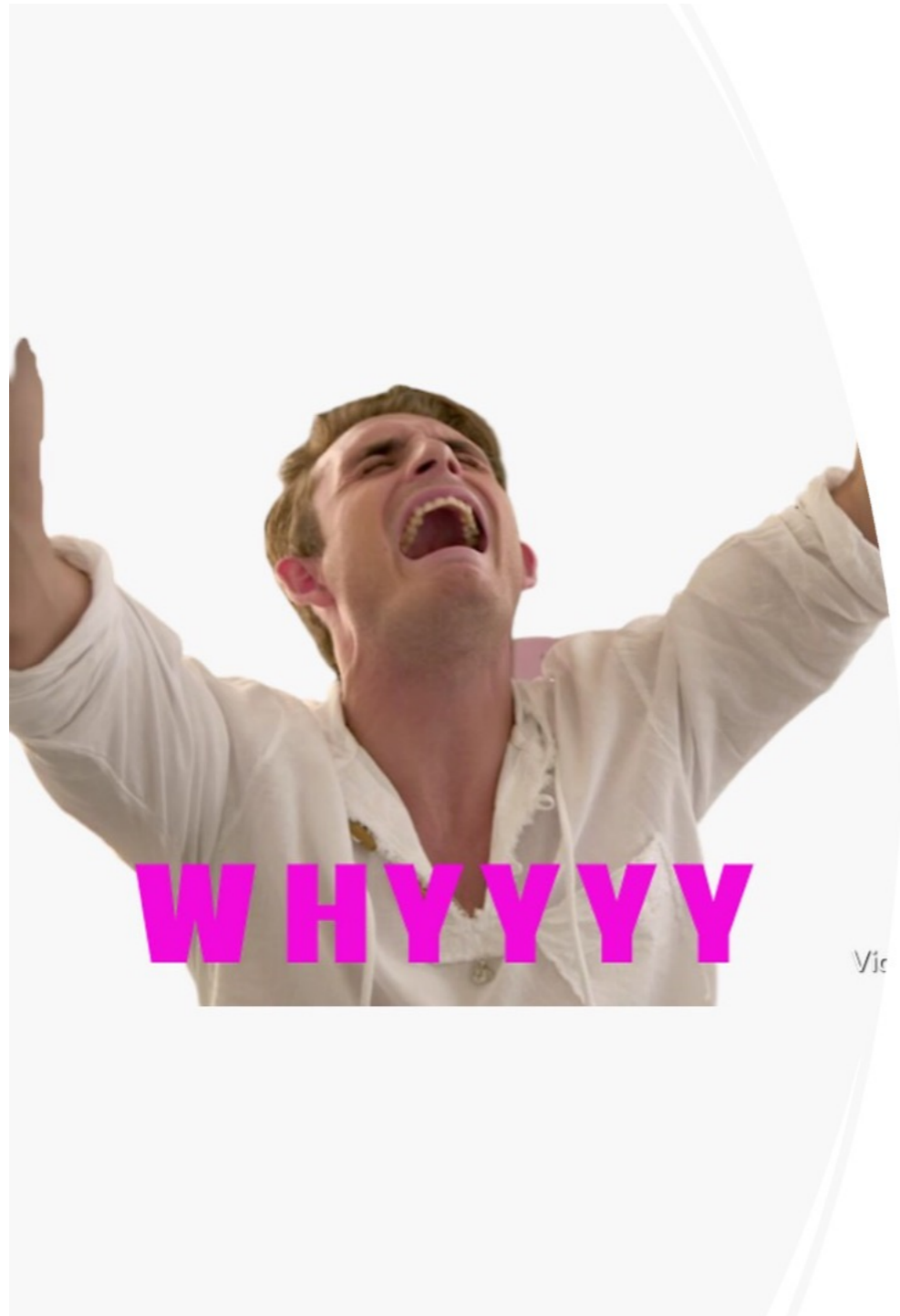
Why shell?



Harvard Chan Bioinformatics Core

Basic Shell Module

February 2023

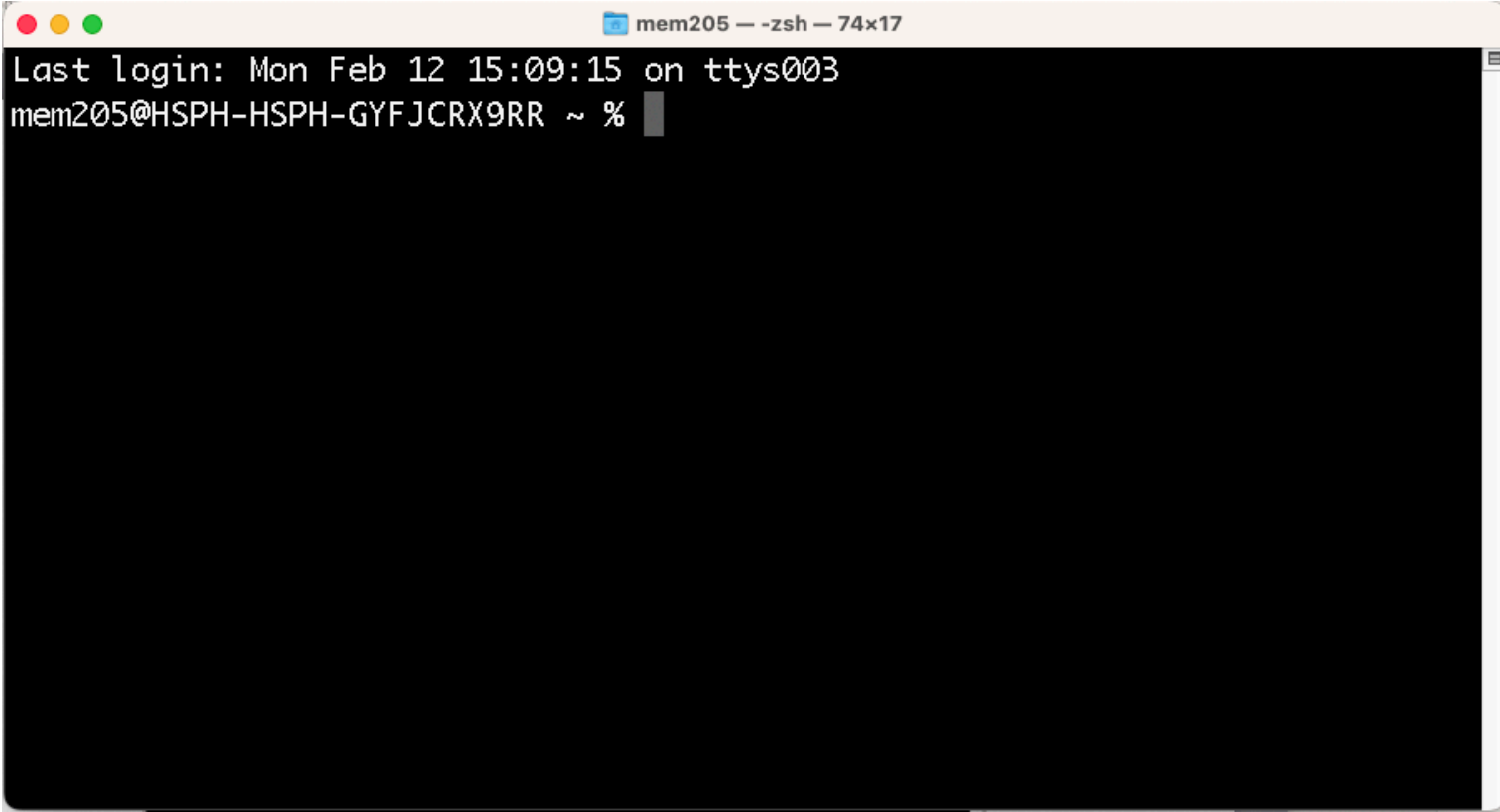


Seriously, why?

*Image source: [VPR sticker by Violetmil](#)*

What is Shell?

Shell - a program that allows users to control  
Unix/Linux OS with text commands

A screenshot of a terminal window. The window title bar shows "mem205 --zsh-- 74x17". The terminal content displays "Last login: Mon Feb 12 15:09:15 on ttys003" followed by the prompt "mem205@HSPH-HSPH-GYFJCRX9RR ~ %" with a cursor. The rest of the terminal area is black.

```
mem205 --zsh-- 74x17
Last login: Mon Feb 12 15:09:15 on ttys003
mem205@HSPH-HSPH-GYFJCRX9RR ~ %
```

# Unix /Linux - The operating systems of High performance computers

---

Unix /Linux - The operating systems of  
High performance computers

---

Shell - a program that allows users to control  
Unix/Linux OS with text commands

---

Unix /Linux - The operating systems of  
High performance computers

---

Shell - a program that allows users to control  
Unix/Linux OS with text commands

---

Bash - the most prevalent kind of shell

# The bottom line

If you plan to process raw high throughput sequencing data yourself, you will need to learn shell.



# 1. You need more resources than what is available on your laptop

- ❖ Sequence data files are LARGE
- ❖ Processing these data require increased CPU and memory
- ❖ High performance compute clusters have the necessary resources!



## 2. Many bioinformatics tools are only available as command-line tools



### 3. Many HTS filetypes are binary.

- ❖ Binary files are not human readable
- ❖ Binary files need an interpreter



*Image source: [Flickr Commons](#)*

# 4. There are many useful commands that can help you work with enormous data files

- ✿ Commands for easily viewing files: less, cat, head, tail
- ✿ More advanced finding and retrieving information and patterns in data with sed, awk and grep

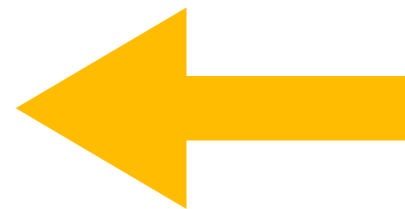
```
0 ##gff-version 3.2.1
1 ##sequence-region ctg123 1 1497228
2 ctg123 . gene 1000 9000 . + . ID=gene00001;Name=EDEN
3 ctg123 . TF_binding_site 1000 1012 . + . ID=tfbs00001;Parent=gene00001
4 ctg123 . mRNA 1050 9000 . + . ID=mRNA00001;Parent=gene00001;Name=EDEN.1
5 ctg123 . mRNA 1050 9000 . + . ID=mRNA00002;Parent=gene00001;Name=EDEN.2
6 ctg123 . mRNA 1300 9000 . + . ID=mRNA00003;Parent=gene00001;Name=EDEN.3
7 ctg123 . exon 1300 1500 . + . ID=exon00001;Parent=mRNA00003
8 ctg123 . exon 1050 1500 . + . ID=exon00002;Parent=mRNA00001,mRNA00002
9 ctg123 . exon 3000 3902 . + . ID=exon00003;Parent=mRNA00001,mRNA00003
10 ctg123 . exon 5000 5500 . + . ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
11 ctg123 . exon 7000 9000 . + . ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
12 ctg123 . CDS 1201 1500 . + 0 ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
13 ctg123 . CDS 3000 3902 . + 0 ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
14 ctg123 . CDS 5000 5500 . + 0 ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
15 ctg123 . CDS 7000 7600 . + 0 ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
16 ctg123 . CDS 1201 1500 . + 0 ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
17 ctg123 . CDS 5000 5500 . + 0 ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
18 ctg123 . CDS 7000 7600 . + 0 ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
19 ctg123 . CDS 3301 3902 . + 0 ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
20 ctg123 . CDS 5000 5500 . + 1 ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
21 ctg123 . CDS 7000 7600 . + 1 ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
22 ctg123 . CDS 3391 3902 . + 0 ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
23 ctg123 . CDS 5000 5500 . + 1 ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
24 ctg123 . CDS 7000 7600 . + 1 ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
```

# 5. Automation is the name of the game

- ❖ Launch many jobs with one command
- ❖ Code is used and reused to iterate tasks over multiple files
- ❖ Parallelization to complete tasks using multiple cores and increase speed!



Image source: [Stock Image](#)



This could be you watching your analysis run.

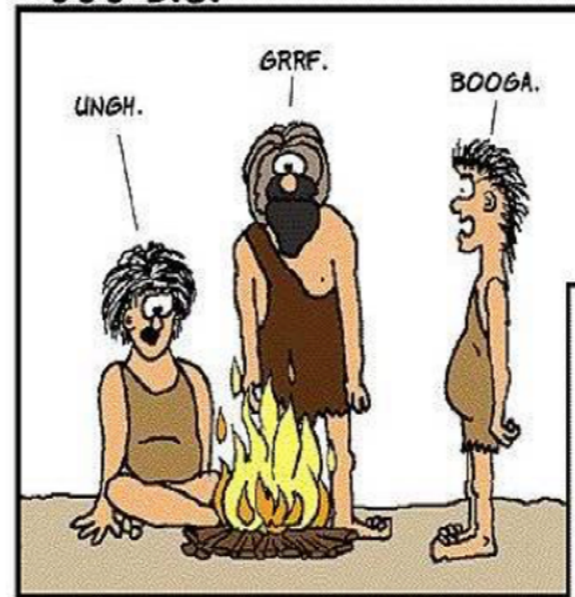
# 6. Bonus! Maybe understand some coding jokes?



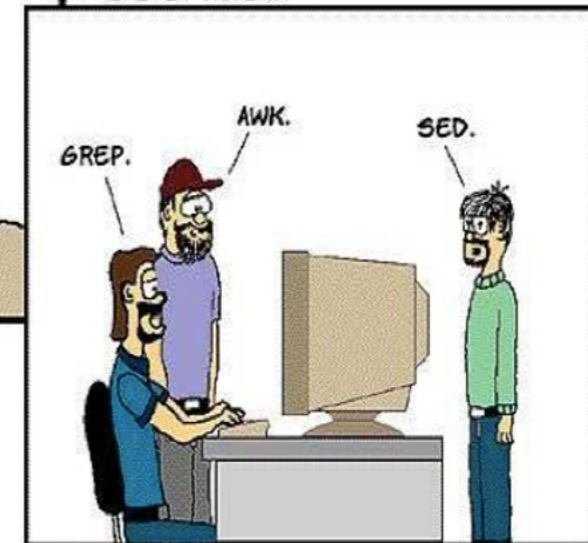
Image source: [xkcd](#)

## EVOLUTION OF LANGUAGE THROUGH THE AGES.

6000 B.C.



2000 A.D.



COPYRIGHT (C) 1999 ILLIAD

[HTTP://WWW.USERFRIENDLY.ORG/](http://www.userfriendly.org/)