Introduction to Bioinformatics

Chris Miller, Ph.D.
Washington University in St. Louis

Bioinformatics Workshop 2025-2026

Supported by – ICTS Precision Health

- We aim to catalyze genomic research by providing grant review, development services, guidance and resources for genomic researchers and genomics education in the community.

Cite the **NIH CTSA Grant #UL1 TR002345** when research is supported by ICTS/CTSA funding or any ICTS Core Services

BFX Workshop – contact John if you haven't received the following

- Slack access, welcome email, Outlook bfx-workshop-2025 group invite

Register for BFX

https://redcap.link/BFX2025

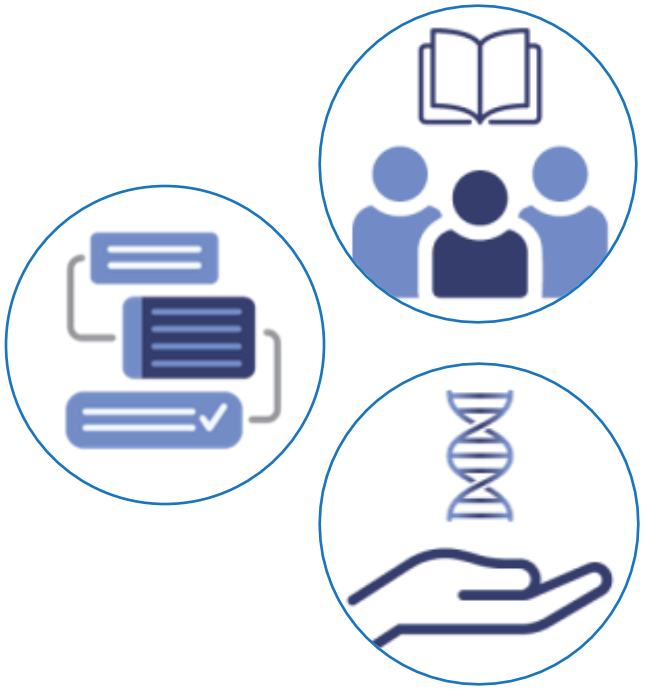
icts-precisionhealth.wustl.edu

j.mckenzie@wustl.edu

Precision Health Led Projects

icts-precisionhealth.wustl.edu
j.mckenzie@wustl.edu

- **Pilot funding & Research reviews**
  - Precision Health Innovation Awards; ICTS Research Development Program



- **Return of Results (ROR) for Research Participants**
  - Genetic counseling, process for returning ACMG secondary results
- **Genomic Database Access and Submission**
  - UK Biobank, All of Us Research Program, dbGaP, AnVIL, SRA
  - Assistance to submit human genomic data to shared repository
- **Institutional Genomic Consent**
  - One Protocol One Consent, BJC-Webb electronic biobank
- **Community Education & Engagement**
  - Precision Health for the Ages Workshop Series

Providing Support For

- **Core Services**
  - WU Biological Therapy Core Facility (BTCF), McDonnell Genome Institute (MGI)
- **Informatics Tools for Precision Health**
  - Bioinformatics Workshop (BFX), pVAC, CIViC,
- **Communications and Outreach**
  - Women in Innovation and Technologies (WIT) program, EQUALIZE program through OTM
- **Educational Opportunities**
  - Precision medicine pathway, Bioinformatics Workshop (BFX), Genomics in Medicine

Institute of Clinical and Translational Sciences


WHO THE HELL ARE YOU?
NETFLIX

Who we are, and why you should trust us

Chris Miller, Ph.D.

Course Director
Associate Professor
Division of Oncology

John Garza

Course Coordinator/TA
Bioinformatics/Genome Analytics
Programmer

20 years of experience in
Bioinformatics and Computational Biology

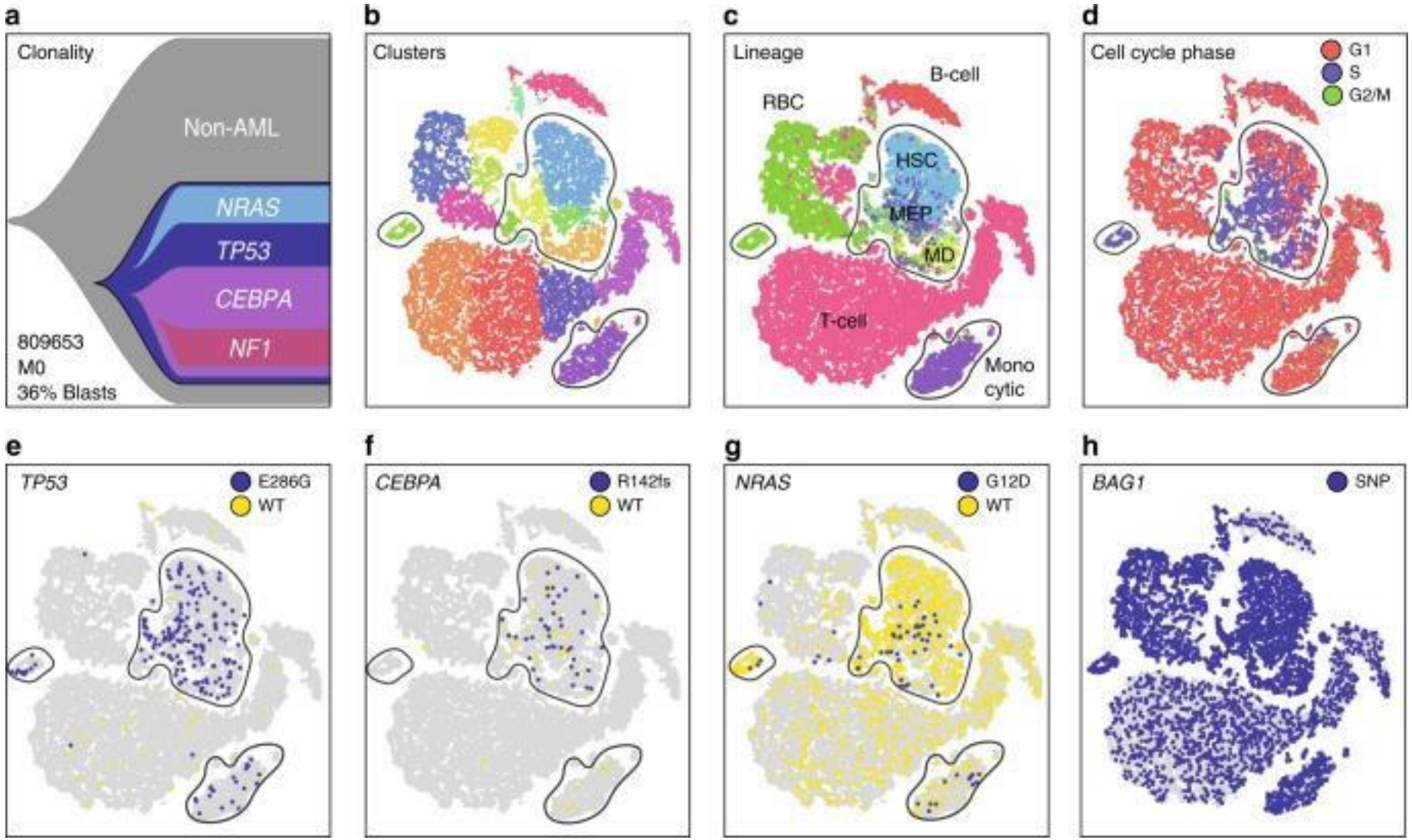
Jenny McKenzie – ICTS
Precision Health Program Scientist

Other Lecturers/Organizers include:

Jason Walker
Obi Griffith
Jennifer Foltz
Juan Macias
Brigida Rusconi

Why learn bioinformatics?

- Biology is now a quantitative discipline - especially genomics


a
Clonality
Non-AML
NRAS
TP53
CEBPA
NF1
809653
M0
36% Blasts
b
Clusters
c
Lineage
B-cell
RBC
HSC
MEP
MD
T-cell
Mono
cytic
d
Cell cycle phase
G1
S
G2/M
e
TP53
E286G
WT
f
CEBPA
R142fs
WT
g
NRAS
G12D
WT
h
BAG1
SNP

Why learn bioinformatics?

- Biology is now a quantitative discipline - especially genomics
- Skills in programming, statistics, and visualization help you get the most out of your data

People who need complex data analysis

People who know how to do complex data analysis

https://hellogiggles.com/news/how-many-people-attend-coachella/
https://www.nytimes.com/2020/07/02/theater/germany-theater-coronavirus.html

Why learn bioinformatics?

- Biology is now a quantitative discipline - especially genomics
- Skills in programming, statistics, and visualization help you get the most out of your data
- We're aiming to teach you the theory and practice of computational biology, with a focus on genomics but lessons that apply broadly

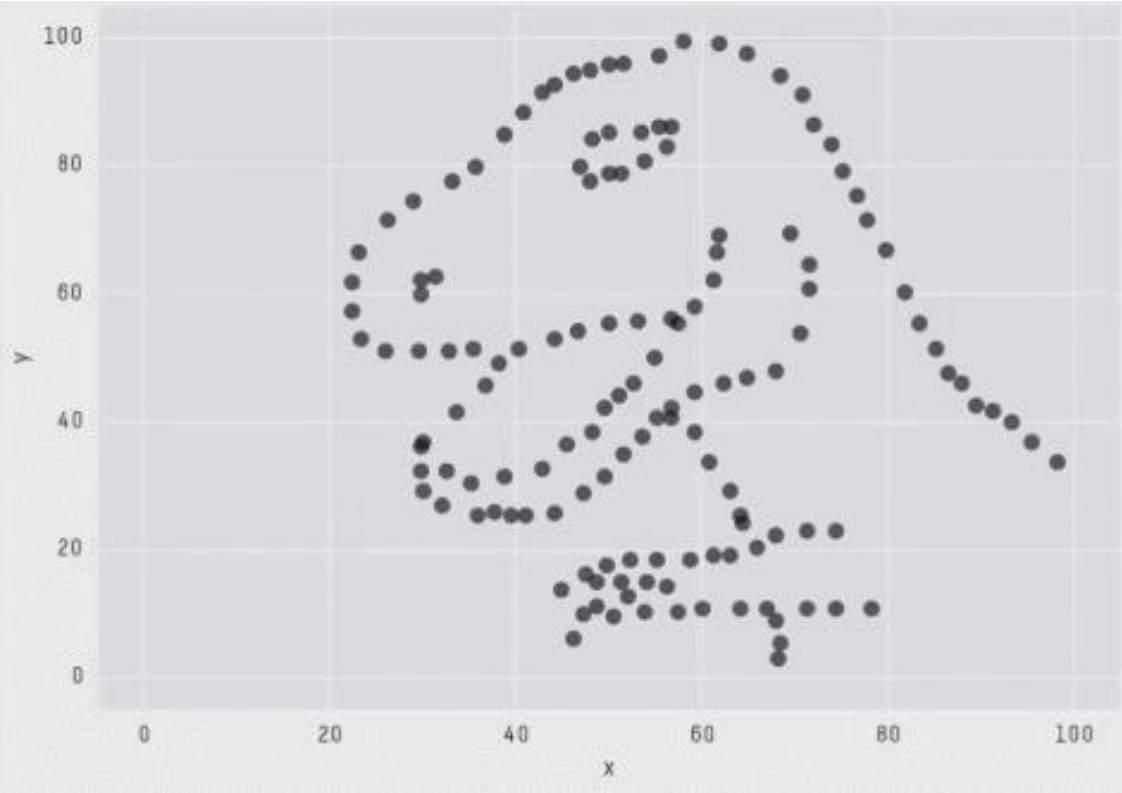
Goals:

- To empower you to improve and expedite your research
- To expose you to new ideas and techniques that may advance your research program

Don’t trust your data

Summary statistics are dangerous

- Visualize your data!
- A picture is worth a thousand p-values



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

Watch out!

- Computational analyses require controls too!
- Look at the data and understand it's limitations!
- Don't assume that the data is clean – prove to yourself that it is!

Expectations:

- Check the prerequisites from week 01. Install the software, be familiar with the unix command line, know how to use docker to launch analyses
  - https://github.com/genome/bfx-workshop/tree/master/lectures/week_01

- Most of you are new to computational analysis – *ask questions*!

- Work hard, follow along, and get your money's worth from this course

- The folks teaching and the TAs all know their stuff, *ask questions*!

Course Structure:

- Weekly lecture introducing topic
- Practical exercise allowing you to apply that knowledge
- https://github.com/genome/bfx-workshop
- ICTS Slack instance: #bfx-workshop channel

- Office hours – 30m before and after each lecture
  - help with homework or help with your own projects

The Unix Shell

What is Unix?

- Family of operating systems (just like Windows)

- Many different "flavors"
  - MacOSX (and iOS)
  - Linux
    - Ubuntu, Debian, RedHat, SElinux, etc…
    - Android

- Nearly every high-performance compute cluster
  - local (RIS compute, others)
  - cloud
  - All of the top 500 "supercomputers" run Linux

Terminals can do things that GUIs can't

- The big event had to be postponed due to COVID and now we have to change every instance of "Apr 2020" to "Oct 2024". Problem is, there's a huge nested set of directories containing over 10,000 files!

- Clicking around in Windows explorer is not going to get the job done

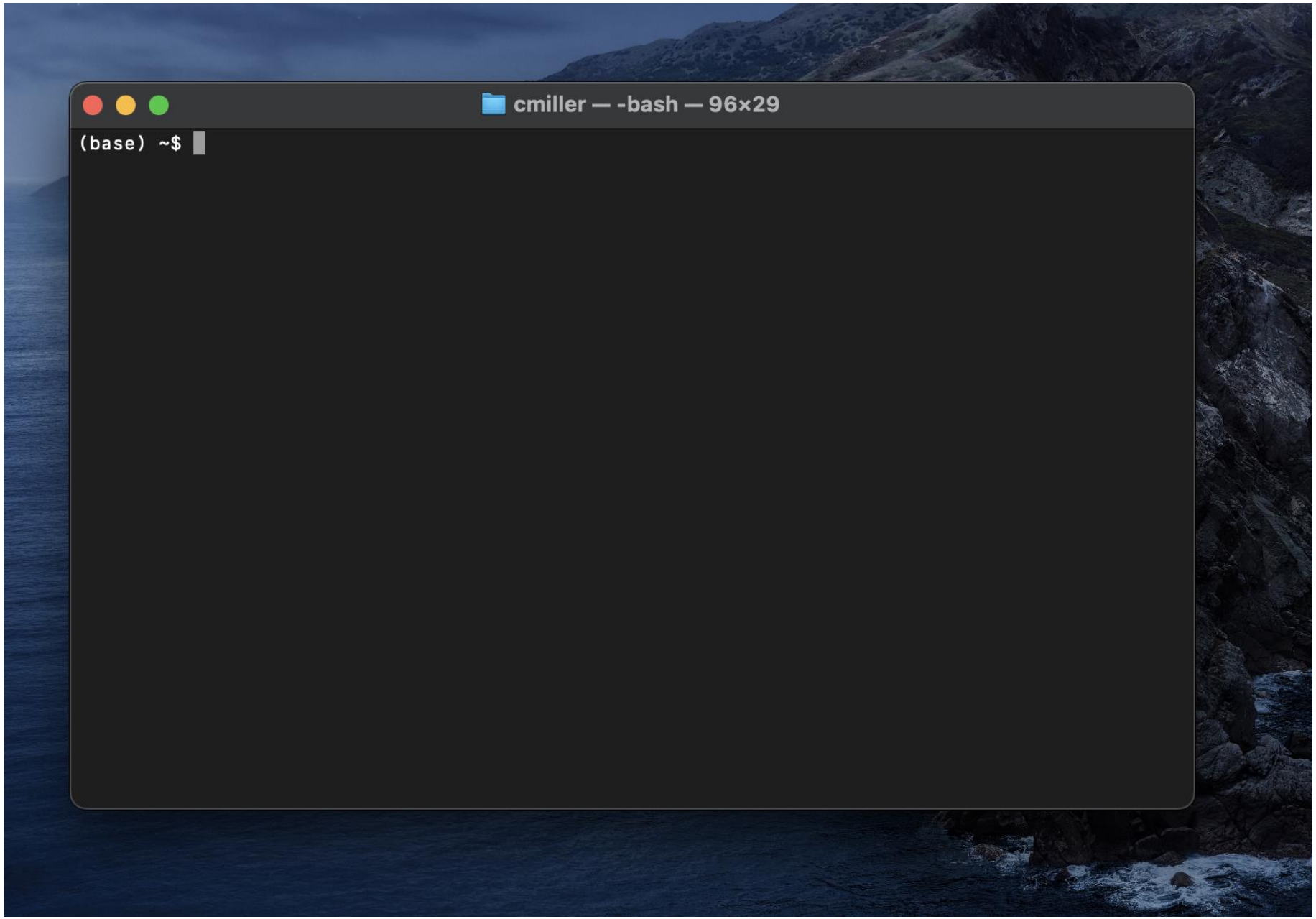
- On a Unix system, that’s just one short line of code:

```
find . -name "*.txt" | xargs -n 1 sed -i.bak 's/Apr 2020/Oct 2024/g'
```

- Seems cryptic at first, but once you learn a little, incredibly powerful!

Unix is the lingua franca of bioinformatics

- high-performance compute clusters run on Unix
- powerful tools for wrangling your data
- writing scripts allows you to do repetitive or error-prone manipulations in a robust and reproducible way
- algorithms for genomics run on the command line


cmiller — -bash — 96×29
(base) ~$

Unix philosophy

- **Modular design**

- Small programs that do one thing well

- Write them to work together

- Handle text streams as the universal interface

https://www.flickr.com/photos/65393447@N00/10711635635

Anatomy of a command

ls -l --color=always ~/workshop

command_name -flag --argument value positional_argument

command

parameters

What options are available?

Generally:

command --help
command -h

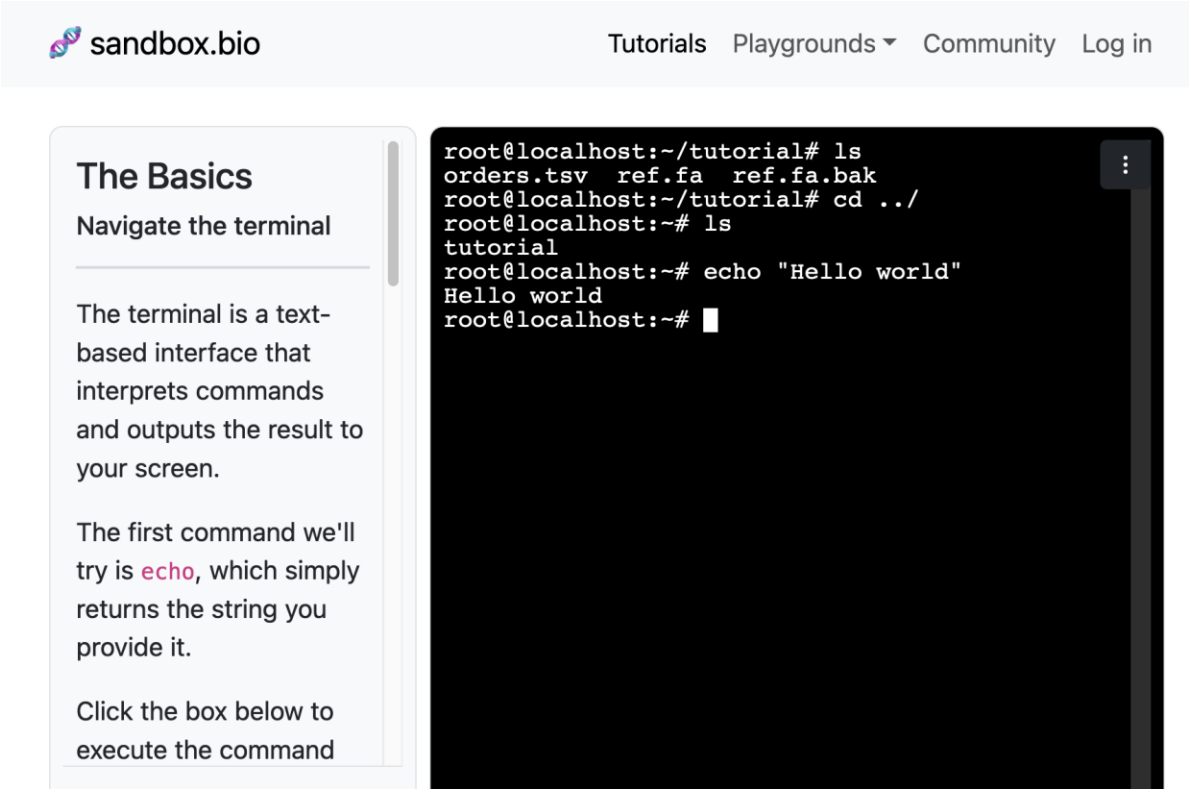
= short help

man command

= manual/more detailed help

Getting started

- ls - list contents of a directory
- echo - print text string
- head/tail – print the first or last lines of a file
- grep – search within a file



https://sandbox.bio/tutorials/terminal-basics/

I'm stuck!

- **Ctrl-C** to interrupt/kill a running process

- **q** quits some interactive commands (e.g. less)

- editing a file with vim?
  - press **Escape**
  - type **:q!**
  - press **Return**

It's not working!

- Did you check case?
  - capital vs lowercase matters!

- Are you in the right directory?
  - use `ls` all the time!

- typos
  - tab-complete is your friend!

Tips and tricks

- Sample and file naming

Good naming for files and directories

- DO use combinations of
  - Alphabetic letters (a-z, A-Z)
  - Numbers (0-9)
  - period (.) underscore (_) and hyphen (-)
- DO be concise, but informative
- DON'T start a filename with a hyphen
  - those are used for parameters
- DON'T use spaces in file names
- DON'T use other special characters

File naming

sample637-bob_mice_w_addback_of_gene_construct_134_plus_gfp_age_3_months.txt

File naming

sample637-bob_mice_w_addback_of_gene_construct_134_plus_gfp_age_3_months.txt

Too long!

Easy to get confused, hard to keep organized

File naming

Mouse Sample A.txt

File naming

Mouse Sample A.txt

Spaces!

```
$ sort Mouse Sample A.txt
sort: No such file or directory
```

looking for files "Mouse" "Sample" and "A.txt"

```
sort "Mouse Sample A.txt"  or  sort Mouse\ Sample\ A.txt
```

would work, but is a pain

File naming

sample123_Tp53+/-_het&Dox*_a.txt

File naming

sample123_Tp53+/-_het&Dox*_a.txt

Special characters!

Unix dirs:
/home/cmiller/workshop/sample123_Tp53+/-_het&Dox*_a.txt

Stick with dashes (-) and underscores(_)
use "plus" or "with" instead of "+"

File naming

-sample123.txt

File naming

-sample123.txt

Starts with hyphen!

```
sort -sample123.txt

sort: invalid option – sample123.txt
```

Sample naming

sample637.tsv
647sample2.tsv
sample983_batch3.tsv

Sample naming

sample637.tsv
647sample2.tsv
sample983_batch3.tsv

sample_637_batch1.tsv
sample_647_batch2.tsv
sample_983_batch3.tsv

Inconsistent!

Sample naming

sample1.tsv
sample2.txt
sample3.tsv
. . .
sample10.tsv
sample11.tsv

Sample naming

sample1.tsv
sample2.txt
sample3.tsv
sample10.tsv
sample11.tsv

```
$ ls
sample1.tsv
sample10.tsv
sample11.tsv
sample2.txt
sample3.tsv
```

Sample naming

sample1.tsv
sample2.txt
sample3.tsv
sample10.tsv
sample11.tsv

```
$ ls
sample1.tsv
sample10.tsv
sample11.tsv
sample2.txt
sample3.tsv
```

```
$ ls
sample01.tsv
sample02.txt
sample03.tsv
sample10.tsv
sample11.tsv
```

Sample naming

sample637
sample647
sample983

Sample naming

sample637	Mouse_637_TP53_KO_WGBS
sample647	Mouse_647_TP53_WT_WGBS
sample983	Mouse_983_TP53_KO_WGBS

Not very informative!

Sample naming

Some good recent examples:

M_RD_57404-CBFB-VavCre_h3k4mono_ChIP
M_RD_57404-CBFB-VavCre_h3k27ac_ChIP

M_YL_NPM9-3A-day1_BM_WGBS
M_YL_NPM9-3L-day1_BM_WGBS

Practice

- https://github.com/genome/bfx-workshop

- Go to lectures/week02/