Bioinformatics Crash Course

Ian Misner Ph.D.

Bioinformatics Coordinator

UMD Bioinformatics Core

The Plan

- Monday
  - Introductions
  - Linux and Python Hands-on Training
- Tuesday
  - NGS Introduction
  - RNAseq with Sailfish (Dr. Steve Mount, CBCB)
  - RNAseq with Tuxedo package
- Wednesday
  - Genome Sequencing Introduction
  - Genome Assembly and QC
  - Metagenomics (Dr. Mihai Pop, CBCB)
- Thursday
  - Genome Annotation
  - PacBio Genome Assembly (Matt Conte)
  - Review Genome Assembly and Annotation
- Friday
  - Cloud computing and Galaxy
    - Variant Detection and RNAseq analysis
- Each day we can have a Q&A session to find out what works or doesn't work as well as try to address any topics we haven't covered.

Break

- http://www.biology.umd.edu/files/biology/bioinformatics/Workshop_July_2014.pdf

UMD Bioinformatics Core

- Mission: To provide users with the bioinformatic services, support, and education necessary to advance their research program.
- The Core has partnered with the Division of IT to provide the necessary computational resources needed for these demanding analyses.

Bioinformatics Core Services

- Raw data processing
- Genome and transcriptome assembly
- RNAseq analysis
- Variant discovery
- Grant writing support
- Experimental design assistance
- Workflow and pipeline construction
- Custom analyses

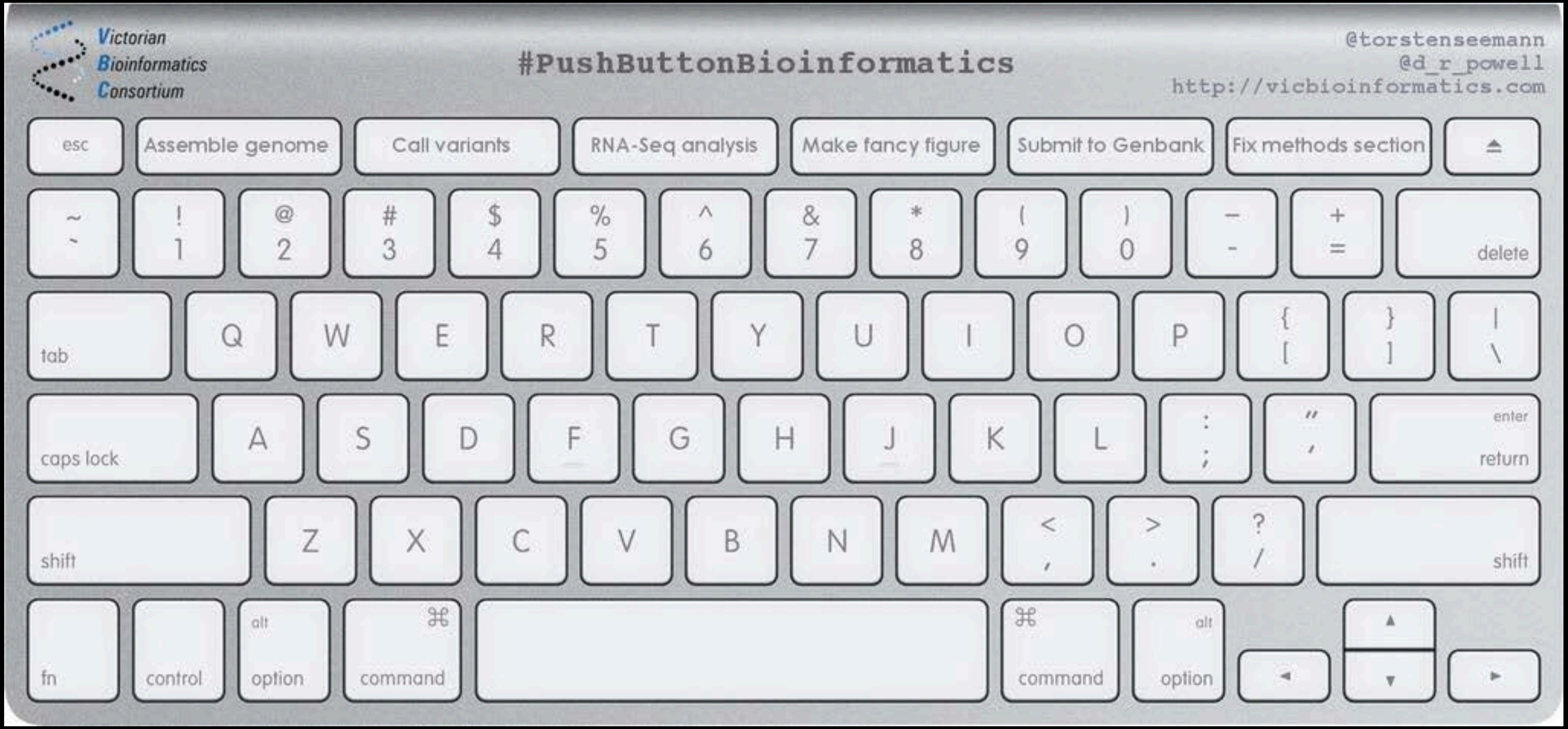
The Goal

- Allow you to conduct your own analysis.
- Get you comfortable using the command line.
- Introduce programing, HPCC's, and DeepThought2.
- Learn some of the best practices with experimental design and data analysis.
- Avoid common pitfalls with data processing.

What is Bioinformatics?

- Interdisciplinary field combining:
  - Computer science
  - Statistics
  - Mathematics
  - And Biology
- Lots of different areas of expertise:
  - Biological programing
  - Software development
  - Hardware development
  - Experimental design
- Difficult for an individual to be an expert in all areas.
- HIGHLY COLLABRATIVE!

What it isn't…

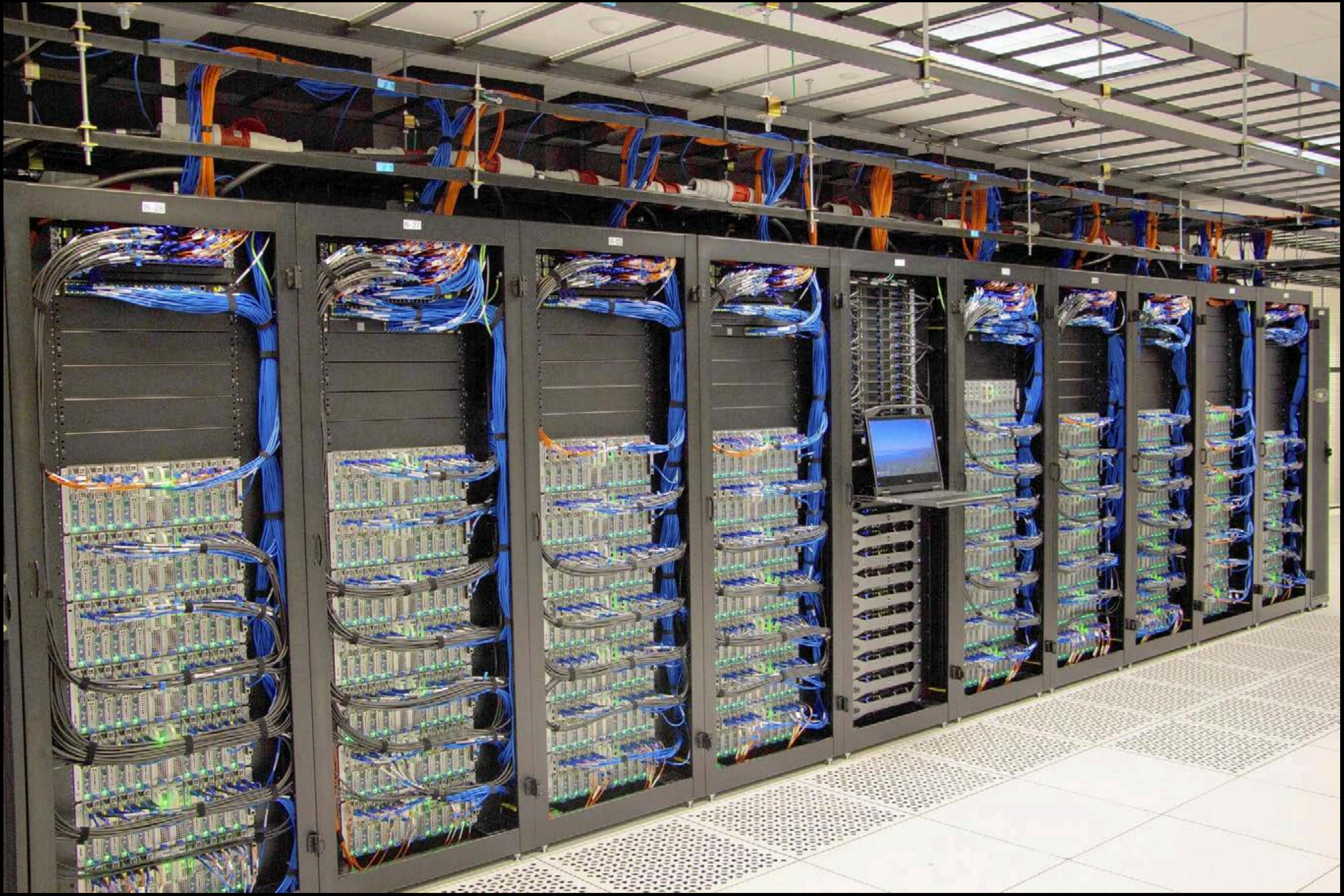

Victorian
Bioinformatics
Consortium
#PushButtonBioinformatics
@torstenseemann
@d_r_powell
http://vicbioinformatics.com
esc
Assemble genome
Call variants
RNA-Seq analysis
Make fancy figure
Submit to Genbank
Fix methods section
~ ` ! 1 @ 2 # 3 $ 4 % 5 ^ 6 & 7 * 8 (9) 0 – - + = delete
tab Q W E R T Y U I O P { [}] | \
caps lock A S D F G H J K L : ; " ' enter return
shift Z X C V B N M < , > . ? / shift
fn control alt option ⌘ command ⌘ command alt option ◂ ▴ ▾ ▸

What is Bioinformatics?

- Bioinformatics is experimental.
- Tools and packages are under constant development and redesign.
- Best practices are only just starting to be determined.
- Always do your own checking … don't assume a program is producing valid information just because there is some output.
- Garbage in … Garbage out!

Tools of the Trade

- Mostly open source tools.
  - These have published code that people can review, modify, and correct.
- This does not necessary mean free.
- Computers, lots of computers.


B-28
B-22
B-25

Tools of the Trade

- Mostly open source tools.
  - These have published code that people can review, modify, and correct.
- This does not necessarily mean free.
- Computers, lots of computers.
- Mountains of Next Generation Sequencing (NGS) Data.

http://www.genomicglossaries.com/images/shenemangenome.gif

Bioinformatic Platforms

- Linux Command Line
  - Python, Perl, R, bash, etc.
- iPlant, iAnimal etc
  - Grant funded, programmer support, intuitional support
- Galaxy
  - Heavy community support and funding.
- Commercial software Geneious, CLC, etc.

Basic File Formats

- FASTA
- FASTQ
- SAM
- BAM

FASTA

```
>My_gene|some description
AGAAAATAGAGAGGCCAGACGATAGATAGAGATCAGCCC
CAGACGCGCGAA
```

- Text based representation of DNA or protein sequence.
- First line starts with a > and is the sequence description.
- The next line is the sequence.
- No standard file extention
  - .fa .fasta .fas

FASTQ

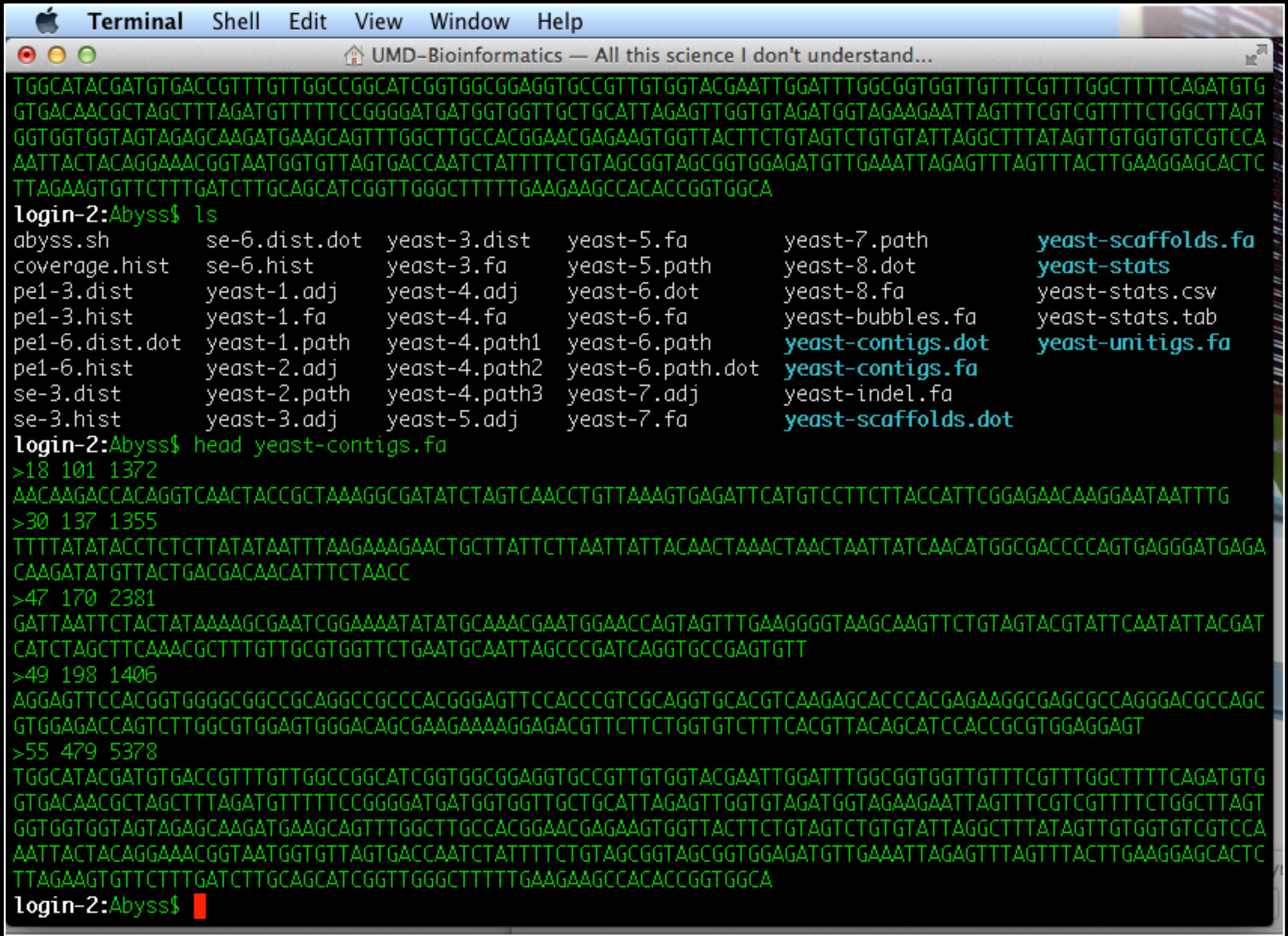
- Fasta with quality information

```
@HWI-EAS225:3:1:2:854#0/1
GGGGGGAAGTCGGCAAAATAGATCCGTAACTTCGGG
+HWI-EAS225:3:1:2:854#0/1
a`abbbbabaabbababb^`[aaa`_N]b^ab^``a
@HWI-EAS225:3:1:2:1595#0/1
GGGAAGATCTCAAAAACAGAAGTAAAACATCGAACG
+HWI-EAS225:3:1:2:1595#0/1
a`abbbababbbabbbbbbabb`aaababab\aa_`
```

Plan

- That covers just the basics.
- Today we are going to work on computer skills
  - Linux
  - Python

Linux



```
TGGCATACGATGTGACCGTTTGTTGGCCGGCATCGGTGGCGGAGGTGCCGTTGTGGTACGAATTGGATTTGGCGGTGGTTGTTTCGTTTGGCTTTTCAGATGTG
GTGACAACGCTAGCTTTAGATGTTTTTCCGGGGATGATGGTGGTTGCTGCATTAGAGTTGGTGTAGATGGTAGAAGAATTAGTTTCGTCGTTTTCTGGCTTAGT
GGTGGTGGTAGTAGAGCAAGATGAAGCAGTTTGGCTTGCCACGGAACGAGAAGTGGTTACTTCTGTAGTCTGTGTATTAGGCTTTATAGTTGTGGTGTCGTCCA
AATTACTACAGGAAACGGTAATGGTGTTAGTGACCAATCTATTTTCTGTAGCGGTAGCGGTGGAGATGTTGAAATTAGAGTTTAGTTTACTTGAAGGAGCACTC
TTAGAAGTGTTCTTTGATCTTGCAGCATCGGTTGGGCTTTTTGAAGAAGCCACACCGGTGGCA
login-2:Abyss$ ls
abyss.sh        se-6.dist.dot  yeast-3.dist   yeast-5.fa        yeast-7.path         yeast-scaffolds.fa
coverage.hist   se-6.hist      yeast-3.fa     yeast-5.path      yeast-8.dot          yeast-stats
pe1-3.dist      yeast-1.adj    yeast-4.adj    yeast-6.dot       yeast-8.fa           yeast-stats.csv
pe1-3.hist      yeast-1.fa     yeast-4.fa     yeast-6.fa        yeast-bubbles.fa     yeast-stats.tab
pe1-6.dist.dot  yeast-1.path   yeast-4.path1  yeast-6.path      yeast-contigs.dot    yeast-unitigs.fa
pe1-6.hist      yeast-2.adj    yeast-4.path2  yeast-6.path.dot  yeast-contigs.fa
se-3.dist       yeast-2.path   yeast-4.path3  yeast-7.adj       yeast-indel.fa
se-3.hist       yeast-3.adj    yeast-5.adj    yeast-7.fa        yeast-scaffolds.dot
login-2:Abyss$ head yeast-contigs.fa
>18 101 1372
AACAAGACCACAGGTCAACTACCGCTAAAGGCGATATCTAGTCAACCTGTTAAAGTGAGATTCATGTCCTTCTTACCATTCGGAGAACAAGGAATAATTTG
>30 137 1355
TTTTATATACCTCTCTTATATAATTTAAGAAAGAACTGCTTATTCTTAATTATTACAACTAAACTAACTAATTATCAACATGGCGACCCCAGTGAGGGATGAGA
CAAGATATGTTACTGACGACAACATTTCTAACC
>47 170 2381
GATTAATTCTACTATAAAAGCGAATCGGAAAATATATGCAAACGAATGGAACCAGTAGTTTGAAGGGGTAAGCAAGTTCTGTAGTACGTATTCAATATTACGAT
CATCTAGCTTCAAACGCTTTGTTGCGTGGTTCTGAATGCAATTAGCCCGATCAGGTGCCGAGTGTT
>49 198 1406
AGGAGTTCCACGGTGGGGCGGCCGCAGGCCGCCCACGGGAGTTCCACCCGTCGCAGGTGCACGTCAAGAGCACCCACGAGAAGGCGAGCGCCAGGGACGCCAGC
GTGGAGACCAGTCTTGGCGTGGAGTGGGACAGCGAAGAAAAGGAGACGTTCTTCTGGTGTCTTTCACGTTACAGCATCCACCGCGTGGAGGAGT
>55 479 5378
TGGCATACGATGTGACCGTTTGTTGGCCGGCATCGGTGGCGGAGGTGCCGTTGTGGTACGAATTGGATTTGGCGGTGGTTGTTTCGTTTGGCTTTTCAGATGTG
GTGACAACGCTAGCTTTAGATGTTTTTCCGGGGATGATGGTGGTTGCTGCATTAGAGTTGGTGTAGATGGTAGAAGAATTAGTTTCGTCGTTTTCTGGCTTAGT
GGTGGTGGTAGTAGAGCAAGATGAAGCAGTTTGGCTTGCCACGGAACGAGAAGTGGTTACTTCTGTAGTCTGTGTATTAGGCTTTATAGTTGTGGTGTCGTCCA
AATTACTACAGGAAACGGTAATGGTGTTAGTGACCAATCTATTTTCTGTAGCGGTAGCGGTGGAGATGTTGAAATTAGAGTTTAGTTTACTTGAAGGAGCACTC
TTAGAAGTGTTCTTTGATCTTGCAGCATCGGTTGGGCTTTTTGAAGAAGCCACACCGGTGGCA
login-2:Abyss$
```

Python

- http://pythonforbiologists.com/