LIVE FREE OR DIE

UNIX

Registered TM of UNIX System Laboratories,
A wholly owned subsidiary of Novell, Inc.
What computers can run Unix?

Apple OS X Macs

Google's Android phones

Wireless internet routers

Airplane entertainment systems
The Terminal Window

the shell, the prompt, the command line
The Terminal Window

Make it comfortable to work in:

- Resize the window
- Change your font size
- Open multiple terminal windows
Obtain a cheat sheet

google “unix commands”
In UNIX everything is a file organized in a hierarchy
/home/catchen/working
Create a series of directories

% mkdir shell
% cd shell
% ls

research
% cd research/
% ls
seq
% cd seq/
% ls

total 8
...
This shell view of the nested directories shell, research, seq, and radtags..... is equivalent to this GUI view of the same directories.

And the radtags directory is uniquely identified by its path:
/home/ubuntu/shell/research/seq/seq/radtags
Absolute and relative paths

How do I get to the Hotel Zlaty Andel?
Absolute and relative paths

How do I get to the Hotel Zlaty Andel?
Relative Path?

/ 
/home 
/usr 
/bin 
... 
cresko 
catchen 
/lib 
/bin 
... 
working 
research 
...
Special files -- ‘dot dot’
../working
Absolute Path:   /home/catchen/working/foo
Relative Path:   ../working/foo
Absolute Path:  /home/catchen/working/foo
Absolute Path:  /home/catchen/research/foo
Relative Path:      ../working/foo
Relative Path:       ./foo
Absolute and relative paths

```
ubuntu@ip-18-4-193-188:~$ mkdir shell
ubuntu@ip-18-4-193-188:~$ cd shell
ubuntu@ip-18-4-193-188:~$ ./shell$ mkdir research
ubuntu@ip-18-4-193-188:~$ ./shell$ ls research
ubuntu@ip-18-4-193-188:~$ ./shell$ cd research/
ubuntu@ip-18-4-193-188:~$ ./shell/research$ mkdir seq
ubuntu@ip-18-4-193-188:~$ ./shell/research$ ls seq
ubuntu@ip-18-4-193-188:~$ ./shell/research/seq$ mkdir radtags
ubuntu@ip-18-4-193-188:~$ ./shell/research/seq$ cd radtags/
ubuntu@ip-18-4-193-188:~$ ./shell/research/seq/radtags$ ls
ubuntu@ip-18-4-193-188:~$ ./shell/research/seq/radtags$ ls -la
```

total 8
drwxrwxr-x 2 ubuntu ubuntu 4096 2012-03-06 23:08 .
drwxrwxr-x 3 ubuntu ubuntu 4096 2012-03-06 23:08 ..

```
ubuntu@ip-18-4-193-188:~$ ./shell/research/seq/radtags$ pwd
/home/ubuntu/shell/research/seq/radtags
ubuntu@ip-18-4-193-188:~$ ./shell/research/seq/radtags$ 
```

Special Files

`dot`

`dot dot`
Binary programs - ls, cp, mkdir, etc.
Relative and absolute paths

A shortcut to your ‘home’, tilde:

~

Moving through the filesystem:

cd

Knowing where you are:

pwd
Relative and absolute paths

/home/tgac/shell/research/seq/radtags

```
% ls .
% ls ..
% ls ../../../
% cd ~/
% cd shell/research
% pwd
```
Are you typing? You’re doing it wrong.

**Tab-completion:**

- Tab once to complete uniquely
- Tab twice to see all possible completions

**Up-arrow:**

- Previous commands can be found by pressing “up-arrow”

'history'

```
% cd /etc
% ls c <tab>
% pwd
% ls c <tab><tab>
```
Three variants to `ls`

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ls -l</strong></td>
<td><strong>ls -la</strong></td>
<td><strong>ls -lh</strong></td>
</tr>
<tr>
<td>provides a <em>long</em> listing</td>
<td>includes <em>all</em> files, even hidden files</td>
<td>displays file sizes in <em>human</em> readable numbers</td>
</tr>
</tbody>
</table>
## Four ways to view a text file

<table>
<thead>
<tr>
<th>more</th>
<th>head</th>
<th>tail</th>
<th>cat</th>
</tr>
</thead>
<tbody>
<tr>
<td>view a text file one screen full at a time</td>
<td>view the top 15 lines of a file</td>
<td>view the last 15 lines of a file</td>
<td>spit the whole file at once</td>
</tr>
<tr>
<td>space-bar: scroll q: quit</td>
<td>-n num controls the number of lines</td>
<td>-n num controls the number of lines</td>
<td></td>
</tr>
</tbody>
</table>
Explore the file hierarchy

1. Move to the directory /etc
   • What is the first line of the file ‘hosts’ in the directory /etc?
   • What is the relative file path to get to /var/log from here?
   • What is the absolute path?

2. Move to the directory /var/log/
   • What is the contents on line 73 of the dpkg.log file?
   • Without changing directories, what is the second line of the cpuinfo file in the proc directory?
     • What is the command to read this file with a relative path?
     • An absolute path?

3. Move back to the root, what directories do you see?

4. Move back home, what are three ways to move home from the root?
Copy example files

Return to the directory in your home called ‘shell’.

**TSV file:**
~/unix/batch_1.genotypes_1.loc.gz

**FASTQ file:**
~/unix/s_1_sequence.txt.gz

**Tar Archive:**
~/unix/samples.tar.gz
What is a tar archive?

\[ \text{tar} = \text{tape archive} \]
Compress / Decompress

gzip / gunzip

batch_1.genotypes_1.loc.gz

s_1_sequence.txt.gz

Gzipped Tar archive

tar xvfz

samples.tar.gz

Tar archive

tar xvf

samples.tar
Sequencing on Illumina’s Flow cell

Source: Illumina Publication 770-2007-002 01May07
Sequencing on Illumina’s Flow cell, ctd.
Sequencing on Illumina’s Flow cell, ctd.

Phred Quality Score

$Q = -10 \log_{10} p$

<table>
<thead>
<tr>
<th>Phred Quality Score</th>
<th>Probability of incorrect base call</th>
<th>Base call accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 in 10</td>
<td>90%</td>
</tr>
<tr>
<td>20</td>
<td>1 in 100</td>
<td>99%</td>
</tr>
<tr>
<td>30</td>
<td>1 in 1000</td>
<td>99.9%</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10000</td>
<td>99.99%</td>
</tr>
<tr>
<td>50</td>
<td>1 in 100000</td>
<td>99.999%</td>
</tr>
</tbody>
</table>

http://en.wikipedia.org/wiki/FASTQ_format
THE FASTQ FILE FORMAT

FASTA

>chromosome7
TTTGTCTGCAGGGGGACACGTCAAAGTCAAACGCAGGCAAGTTTGTGTTTATGTCCAGTGGATCTTTTGATTTT
ACATACTGCAGGGGTAGGATTATCTCTCCTGCAAGGTAACGCCTGCTGTAACCGTTGTTCTTCATCCTTTTT
CCTAACTGCAGGGGCTGTCTTGTCAGGTCTGACAAGACATATGCAGGCTCAATTTTGAGATAATTTGCTCAATATA

FASTQ

@Sequence_137
TTTGTCTGCAGGGGGACACGTCAAAGTCAAACGCAGGCAAGTTTGTGTTTATGTCCAGTGGATCTTTTGATTTT
+Sequence_137
<?@DDDDDHFFHFFBB@GGIACFHGGHGHGGCDHBEAHACHI=@CH.=7ACAHHADECDBCC66(6>@C>5@CACCA

@HWI-ST0747:162:C03AJACXX:3:1108:19763:106771 1:N:0:
TTTGTCTGCAGGGGGACACGTCAAAGTCAAACGCAGGCAAGTTTGTGTTTATGTCCAGTGGATCTTTTGATTTT
+  
<?@DDDDDHFFHFFBB@GGIACFHGGHGHGGCDHBEAHACHI=@CH.=7ACAHHADECDBCC66(6>@C>5@CACCA
### ASCII Code

<table>
<thead>
<tr>
<th>Code</th>
<th>Character</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>&lt;NUL&gt;</td>
</tr>
<tr>
<td>1</td>
<td>&lt;SOH&gt;</td>
</tr>
<tr>
<td>2</td>
<td>&lt;STX&gt;</td>
</tr>
<tr>
<td>3</td>
<td>&lt;ETX&gt;</td>
</tr>
<tr>
<td>4</td>
<td>&lt;EOT&gt;</td>
</tr>
<tr>
<td>5</td>
<td>&lt;ENQ&gt;</td>
</tr>
<tr>
<td>6</td>
<td>&lt;ACK&gt;</td>
</tr>
<tr>
<td>7</td>
<td>&lt;BEL&gt;</td>
</tr>
<tr>
<td>8</td>
<td>&lt;BS&gt;</td>
</tr>
<tr>
<td>9</td>
<td>&lt;TAB&gt;</td>
</tr>
<tr>
<td>10</td>
<td>&lt;LF&gt;</td>
</tr>
<tr>
<td>11</td>
<td>&lt;VT&gt;</td>
</tr>
<tr>
<td>12</td>
<td>&lt;FF&gt;</td>
</tr>
<tr>
<td>13</td>
<td>&lt;CR&gt;</td>
</tr>
<tr>
<td>14</td>
<td>&lt;FS&gt;</td>
</tr>
<tr>
<td>15</td>
<td>&lt;SUB&gt;</td>
</tr>
<tr>
<td>16</td>
<td>&lt;ESC&gt;</td>
</tr>
<tr>
<td>17</td>
<td>&lt;FS&gt;</td>
</tr>
<tr>
<td>18</td>
<td>&lt;FS&gt;</td>
</tr>
<tr>
<td>19</td>
<td>&lt;FS&gt;</td>
</tr>
<tr>
<td>20</td>
<td>&lt;FS&gt;</td>
</tr>
<tr>
<td>21</td>
<td>&lt;FS&gt;</td>
</tr>
<tr>
<td>22</td>
<td>&lt;FS&gt;</td>
</tr>
<tr>
<td>23</td>
<td>&lt;FS&gt;</td>
</tr>
<tr>
<td>24</td>
<td>&lt;FS&gt;</td>
</tr>
<tr>
<td>25</td>
<td>&lt;FS&gt;</td>
</tr>
<tr>
<td>26</td>
<td>&lt;FS&gt;</td>
</tr>
<tr>
<td>27</td>
<td>&lt;FS&gt;</td>
</tr>
<tr>
<td>28</td>
<td>&lt;FS&gt;</td>
</tr>
<tr>
<td>29</td>
<td>&lt;FS&gt;</td>
</tr>
<tr>
<td>30</td>
<td>&lt;FS&gt;</td>
</tr>
<tr>
<td>31</td>
<td>&lt;FS&gt;</td>
</tr>
</tbody>
</table>

8 bits = $2^8$ combinations = 256

```
0x2^7 + 1x2^6 + 1x2^5 + 1x2^4 + 1x2^3 + 0x2^2 + 0x2^1 + 1x2^0 = 121 = y
```

```
1x10^2 + 2x10^1 + 1x10^0 = 121 = y
```
The FASTQ File Format, ctd

@HWI-ST0747:162:C03AJACXX:3:1108:19763:106771 1:N:0:
TTTGTCCTGCAGGGGAGCAGTCATCAGTCAGGATGGATCTTTGATT
+
<?@DDDDDFHHFHFBB@GGIACFHGGHBGHDHBEAHACHI=@CH.=7ACAHHADECDBCC66(6>@C>5@CACCA

Quality Scores

<table>
<thead>
<tr>
<th>ASCII values</th>
<th>S - Sanger Phred+33, raw reads typically (0, 40)</th>
<th>X - Solexa Solexa+64, raw reads typically (-5, 40)</th>
<th>I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)</th>
<th>J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>59</td>
<td>64</td>
<td>73</td>
<td>104</td>
</tr>
</tbody>
</table>

S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 40)
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)

ASCII values 33 - 73 = 0 - 40

‘F’ = 70

70 - 33 = 37

http://en.wikipedia.org/wiki/FASTQ_format
The FASTQ File Format, ctd

@HWI-ST0747:162:C03AJACXX:3:1108:19763:106771 1:N:0:
TTTGCTCTGCAGGGGGACACGTCAAAAGTCAACGCAGGCAAGTTTGTGTTATGTCCAGTGATCTTTTGATTTT
+
<?@DDDDDHFFHFB@GGIACFHGGHBGHGCDHBEAHACHI=@CH.=7ACAHHADECDBCC66(6>@C>5@CACCA

70 - 33 = 37

<table>
<thead>
<tr>
<th>Phred Quality Score</th>
<th>Probability of incorrect base call</th>
<th>Base call accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 in 10</td>
<td>90%</td>
</tr>
<tr>
<td>20</td>
<td>1 in 100</td>
<td>99%</td>
</tr>
<tr>
<td>30</td>
<td>1 in 1000</td>
<td>99.9%</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10000</td>
<td>99.99%</td>
</tr>
<tr>
<td>50</td>
<td>1 in 100000</td>
<td>99.999%</td>
</tr>
</tbody>
</table>

http://en.wikipedia.org/wiki/Phred_quality_score
Count raw reads:

```
wc -l s_1_sequence.txt

grep "@" s_1_sequence.txt
grep -c "@" s_1_sequence.txt

grep -v "@" s_1_sequence.txt
grep -v -c "@" s_1_sequence.txt
```

Count reads with barcode:

```
grep -c "^CGATA" s_1_sequence.txt
```
Special Files

STDIN, STDOUT, STDERR

The Shell’s Killer App: **Pipes**
The Shell’s Killer App: **Pipes**, ctd.
The Shell’s Killer App: **Pipes**, ctd.

So what is the purpose of the program **cat**?
cut

cut \(-f\ 10\) `batch_1.genotypes_1.loc`

cut, capture the output

cut \(-f\ 1-10\) `batch_1.genotypes_1.loc` > genos

cut, pipe the output to grep

cut \(-f\ 2\) `batch_1.genotypes_1.loc` | grep \(-c\) "nnxnp"

cut \(-f\ 1-10,15,17\) `batch_1.genotypes_1.loc` | grep "nnxnp" > genos2

Examine a marker, translating the output

cat `batch_1.genotypes_1.loc` | tr " " "," | grep "^96053"
Decompress the file

1. Count the number of raw reads (250,000)
2. Count the number of reads with barcode CGATA (19,501)
3. Capture all FASTQ records for ACCAT into a file called sample_01.fq (you should get 18352 records, 73408 lines)
4. Determine the count of all barcodes in the file

```plaintext
286 CTAGT
7900 TCAGA
10659 ACTGC
10931 TGACC
11536 GAGAT
11871 CTGAA
14409 CGGCC
14508 TGGTT
18226 GAAAGC
18352 ACCAT
18375 TCGAG
19501 CGATA
23012 AATTT
26336 GCATT
31136 CTAGG
```

1. Use `head` when building a command, `cat` once the command is working
2. Look at the `-n` option for the `head` command, the `-l` option for `wc`
3. The “^” character means “must occur at beginning of line” in a grep search
4. Look at the `grep` options: `-c`, `-v`, `-A`, `-B`
5. Read the man pages for `sort` and `uniq` to learn how to combine them
### Problem Set #1

#### Danger Is. #04 vs Middleton Is. #16 Fst

<table>
<thead>
<tr>
<th>#</th>
<th>Batch ID</th>
<th>Locus ID</th>
<th>Pop 1 ID</th>
<th>Pop 2 ID</th>
<th>Chr</th>
<th>BP</th>
<th>Column</th>
<th>Overall Pi</th>
<th>Fst</th>
<th>Fisher's P</th>
<th>Odds Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>7894</td>
<td>2</td>
<td>3</td>
<td>groupI</td>
<td>11832</td>
<td>19</td>
<td>0.428182</td>
<td>-0.0076252913</td>
<td>0.191294</td>
<td>0.687192</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>7896</td>
<td>2</td>
<td>3</td>
<td>groupI</td>
<td>11900</td>
<td>83</td>
<td>0.328622</td>
<td>0.1775694587</td>
<td>4.35747e-08</td>
<td>5.44667</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>9611</td>
<td>2</td>
<td>3</td>
<td>groupI</td>
<td>49756</td>
<td>48</td>
<td>0.090426</td>
<td>-0.1127451906</td>
<td>0.00109115</td>
<td>0.0898072</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>9611</td>
<td>2</td>
<td>3</td>
<td>groupI</td>
<td>49765</td>
<td>57</td>
<td>0.0132887</td>
<td>-0.1522407447</td>
<td>0.518395</td>
<td>0.493113</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>9611</td>
<td>2</td>
<td>3</td>
<td>groupI</td>
<td>49766</td>
<td>58</td>
<td>0.0133776</td>
<td>-0.1563725438</td>
<td>0.520033</td>
<td>0.501401</td>
<td></td>
</tr>
</tbody>
</table>

...