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
Paths

/home/catchen/working
Create a series of directories

% mkdir shell
% cd shell
% ls

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

total 8

% pwd
/home/ubuntu/shell/research/seq/radtags
%
Paths, cont

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/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?*
Absolute Path

/home/catchen/working
Relative Path?
Special files -- ‘dot’
Special files -- ‘dot dot’
Relative Path

./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

```
% ls .
% ls ..
% ls ../../../
```

Special Files

- `dot`
- `dot dot`
Binary programs - `ls`, `cp`, `mkdir`, etc.

```bash
% ls /bin
```
Relative and absolute paths

A shortcut to your ‘home’, tilde:

~

Moving through the filesystem:

cd

Knowing where you are:

pwd

% ls ~/
% cd ~/
% cd
% 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`

<table>
<thead>
<tr>
<th>Command</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>ls -l</code></td>
<td>Provides a long listing</td>
</tr>
<tr>
<td><code>ls -la</code></td>
<td>Includes all files, even hidden files</td>
</tr>
<tr>
<td><code>ls -lh</code></td>
<td>Displays file sizes in human readable numbers</td>
</tr>
</tbody>
</table>

- `ls -l` provides a long listing.
- `ls -la` includes all files, even hidden files.
- `ls -lh` displays file sizes in human readable numbers.
# 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</td>
<td>view the top 15 lines of</td>
<td>view the last 15 lines of</td>
<td>spit the whole file at</td>
</tr>
<tr>
<td>screen full at a time</td>
<td>a file</td>
<td>a file</td>
<td>once</td>
</tr>
<tr>
<td>space-bar: scroll q:</td>
<td>-n num controls the</td>
<td>-n num controls the</td>
<td></td>
</tr>
<tr>
<td>q: quit</td>
<td>number of lines</td>
<td>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 dmesg 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:**

~/workshop_materials/unix/batch_1.genotypes_1.loc.gz

**FASTQ file:**

~/workshop_materials/unix/s_1_sequence.txt.gz

**Tar Archive:**

~/workshop_materials/unix/samples.tar.gz
What is a tar archive?

tar = 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
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>
The FASTQ File Format

FASTA
>chromosome7
TTTGCTCTGACAGGGGACACGTCAAAGTCAAAGCAGGCAAGTTTGTGTTTATGTCCAGGTGATCTTTTGATTTT
ACATACTGCAGGGTCAGGAGGATTATCTCTCTCTGCAAGGTAACGCCTGCTGTAACGGGTTGTTTCATCATCTTTTT
CCTAAGCTGACGGCCTGCTTTGTCAGGTCTGACAAGAGACATCATAGCAGGGCTCAATTTGAGATAATTGCTCAATATA

FASTQ
@Sequence_137
TTTGCTCTGACAGGGGACACGTCAAAGTCAAAGCAGGCAAGTTTGTGTTTATGTCCAGGTGATCTTTTGATTTT
+Sequence_137
<?@DDDDDFHFFHFBBABGGIACFHGGBHGBHCDHBEAHACHI=@CH.=7ACAHHADECDBCC66(6>@C>5@CAGCA
## ASCII Code

<p>| | | | | | | | | | | | | | | | | | | | | | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>&lt;NUL&gt;</td>
<td>32</td>
<td>!</td>
<td>64</td>
<td>@</td>
<td>96</td>
<td>`</td>
<td>128</td>
<td>Å</td>
<td>160</td>
<td>†</td>
<td>192</td>
<td>č</td>
<td>224</td>
<td>‡</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>&lt;SOH&gt;</td>
<td>33</td>
<td>&quot;</td>
<td>65</td>
<td>A</td>
<td>97</td>
<td>a</td>
<td>129</td>
<td>Å</td>
<td>161</td>
<td>°</td>
<td>193</td>
<td>i</td>
<td>225</td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>&lt;STX&gt;</td>
<td>34</td>
<td>&quot;</td>
<td>66</td>
<td>B</td>
<td>98</td>
<td>b</td>
<td>130</td>
<td>Ç</td>
<td>162</td>
<td>ø</td>
<td>194</td>
<td>ñ</td>
<td>226</td>
<td>,</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>&lt;ETX&gt;</td>
<td>35</td>
<td>#</td>
<td>67</td>
<td>C</td>
<td>99</td>
<td>c</td>
<td>131</td>
<td>É</td>
<td>163</td>
<td>£</td>
<td>195</td>
<td>√</td>
<td>227</td>
<td>&quot;</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
| 4   | <EOT> | 36 | $   | 68 | D   | 100 | d   | 132 | Ñ   | 164 | §   | 196 | f   | 228 | %
| 5   | <ENQ> | 37 | %   | 69 | E   | 101 | e   | 133 | Ö   | 165 | •   | 197 | ≈   | 229 | Å   |
| 6   | <ACK> | 38 | &   | 70 | F   | 102 | f   | 134 | Ü   | 166 | ¶   | 198 | Δ   | 230 | É   |
| 7   | <BEL> | 39 | '   | 71 | G   | 103 | g   | 135 | á   | 167 | ß   | 199 | «   | 231 | À   |
| 8   | <BS>  | 40 | (   | 72 | H   | 104 | h   | 136 | à   | 168 | @   | 200 | »   | 232 | É   |
| 9   | <TAB> | 41 | )   | 73 | I   | 105 | i   | 137 | â   | 169 | ©   | 201 | …   | 233 | É   |
| 10  | <LF>  | 42 | *   | 74 | J   | 106 | j   | 138 | ä   | 170 | ™   | 202 | 234 | í   |
| 11  | <VT>  | 43 | +   | 75 | K   | 107 | k   | 139 | ä   | 171 | '   | 203 | À   | 235 | Í   |
| 12  | <FF>  | 44 | ,   | 76 | L   | 108 | l   | 140 | â   | 172 | "   | 204 | À   | 236 | Í   |
| 13  | <CR>  | 45 | -   | 77 | M   | 109 | m   | 141 | ç   | 173 | ℓ   | 205 | Ö   | 237 | i   |
| 14  | <SO>  | 46 | .   | 78 | N   | 110 | n   | 142 | é   | 174 | Æ   | 206 | CE  | 238 | Ó   |
| 15  | <SI>  | 47 | /   | 79 | O   | 111 | o   | 143 | è   | 175 | Ø   | 207 | æ   | 239 | Ô   |
| 16  | <DE>  | 48 | 0   | 80 | P   | 112 | p   | 144 | é   | 176 | ∞   | 208 | —   | 240 | ¬   |
| 17  | <DC1> | 49 | 1   | 81 | Q   | 113 | q   | 145 | é   | 177 | ±   | 209 | —   | 241 | Ö   |
| 18  | <DC2> | 50 | 2   | 82 | R   | 114 | r   | 146 | í   | 178 | ≤   | 210 | “   | 242 | Ú   |
| 19  | <DC3> | 51 | 3   | 83 | S   | 115 | s   | 147 | ì   | 179 | ≥   | 211 | ”   | 243 | Ú   |
| 20  | <DC4> | 52 | 4   | 84 | T   | 116 | t   | 148 | ì   | 180 | ¥   | 212 | "   | 244 | Ù   |
| 21  | <NAK> | 53 | 5   | 85 | U   | 117 | u   | 149 | í   | 181 | µ   | 213 | '   | 245 | Ñ   |
| 22  | <SYN> | 54 | 6   | 86 | V   | 118 | v   | 150 | ŋ   | 182 | ò   | 214 | ÷   | 246 | —   |
| 23  | <ETB> | 55 | 7   | 87 | W   | 119 | w   | 151 | ó   | 183 | Σ   | 215 | ø   | 247 | —   |
| 24  | <CAN> | 56 | 8   | 88 | X   | 120 | x   | 152 | ó   | 184 | Π   | 216 | Ý   | 248 | —   |
| 25  | <EM>  | 57 | 9   | 89 | Y   | 121 | y   | 153 | ó   | 185 | n   | 217 | Ý   | 249 | —   |
| 26  | <SUB> | 58 | :   | 90 | Z   | 122 | z   | 154 | ó   | 187 | a   | 219 | C   | 251 |
| 27  | <ESC> | 59 | \   | 91 | [   | 123 | {   | 155 | ó   | 187 | ₃   | 219 | C   | 251 |
| 28  | <FS>  | 60 | <   | 92 | \   | 124 | {   | 156 | ú   | 188 | ö   | 220 | <   | 252 | |
| 29  | <GS>  | 61 | =   | 93 | ]   | 125 | }   | 157 | ú   | 189 | Ω   | 221 | >   | 253 | |
| 30  | <RS>  | 62 | >   | 94 | ~   | 126 | ~   | 158 | ú   | 190 | æ   | 222 | fi  | 254 | |
| 31  | <US>  | 63 | ?   | 95 | _   | 127 | <DEL> | 159 | ù   | 191 | ø   | 223 | fl  | 255 | |

8 bits = 2<sup>8</sup> combinations = 256

\[
0 \times 2^7 + 1 \times 2^6 + 1 \times 2^5 + 1 \times 2^4 + 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 = 121 = y
\]

\[
1 \times 10^2 + 2 \times 10^1 + 1 \times 10^0 = 121 = y
\]
The FASTQ File Format, ctd

@HWI-ST0747:162:C03AJACXX:3:1108:19763:106771 1:N:0: TTTGTCTGCAGGGGACACGTCAAAGTCAAACGCAGGCAAGTTTGTGTTTATGTCCAGTGATCTTTTGATT    + C?@DDDDDHFHHFBB@GGIACFHHGHGHHGCDHBEAHACHI=@CH.=7ACAHHADECDDBCC66(6>@C>5@CACCA

### Quality Scores

<table>
<thead>
<tr>
<th>ASCII values</th>
<th>Sanger</th>
<th>Phred+33, raw reads typically (0, 40)</th>
<th>Solexa</th>
<th>Phred+64, raw reads typically (-5, 40)</th>
<th>Illumina 1.3+ Phred+64, raw reads typically (0, 40)</th>
<th>Illumina 1.5+ Phred+64, raw reads typically (3, 40)</th>
<th>Read Segment Quality Control Indicator (bold)</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>S</td>
<td>59</td>
<td>X</td>
<td>64</td>
<td>I</td>
<td>J</td>
<td>L</td>
</tr>
<tr>
<td>59</td>
<td></td>
<td></td>
<td>64</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>104</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>126</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></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

The FASTQ File Format, ctd

@HWI-ST0747:162:C03AJACXX:3:1108:19763:106771 1:N:0:
TTTGTCTGCAGGGGGACACGTCAAAGTCAAACGCAGGCAAGTTTGTGTTTATGTCCAGTGATCTTTTGATT
+
<?@DDDDDHFFHBGB@GGIACFHGGHBGHGCDHBEAHACHI=@CH.=7ACAHHADECDDBCC66 (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.
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

   286 CTAGT
   7900 TCAGA
   10659 ACTGC
   10931 TGACC
   11536 GAGAT
   11871 CTGAA
   14409 CGGCG
   14508 TGGTT
   18226 GAAGC
   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

<table>
<thead>
<tr>
<th>#</th>
<th>Batch ID</th>
<th>Locus ID</th>
<th>Pop 1</th>
<th>Pop 2</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>

...