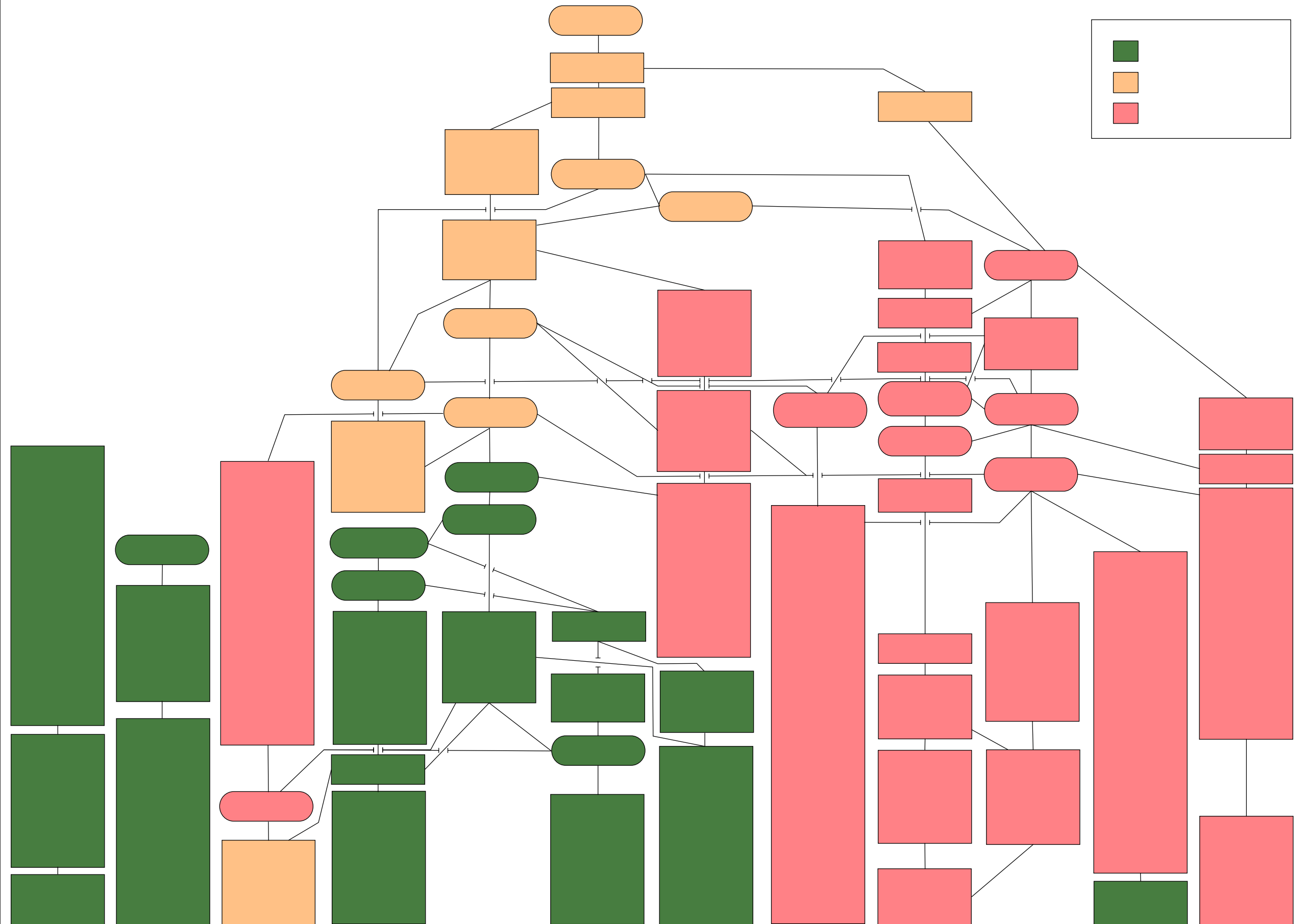


Unix History

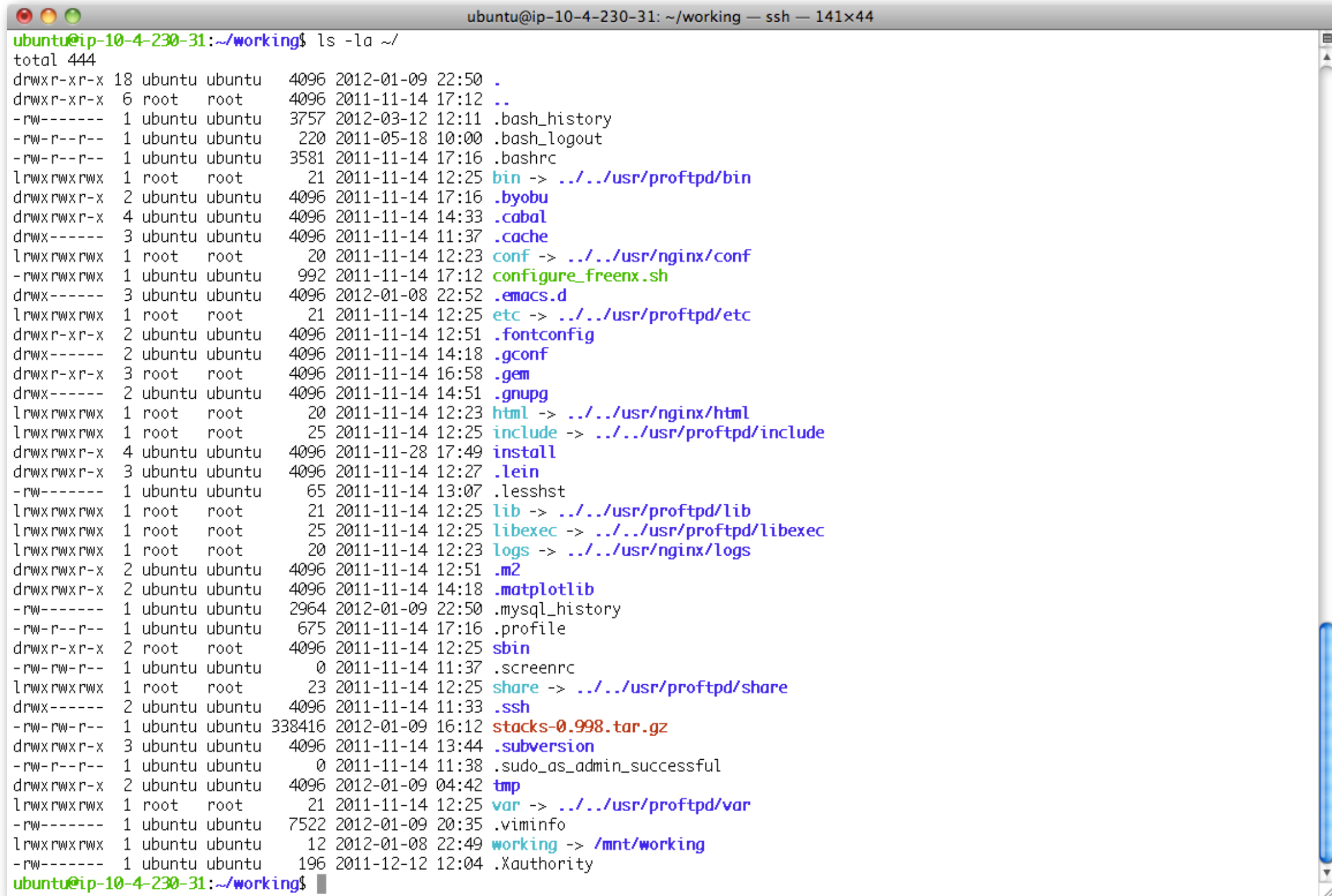


What computers can run Unix?



- Apple OS X Macs
- Google's Android phones
- Most airplane entertainment systems
- Wireless internet routers

The Terminal Window



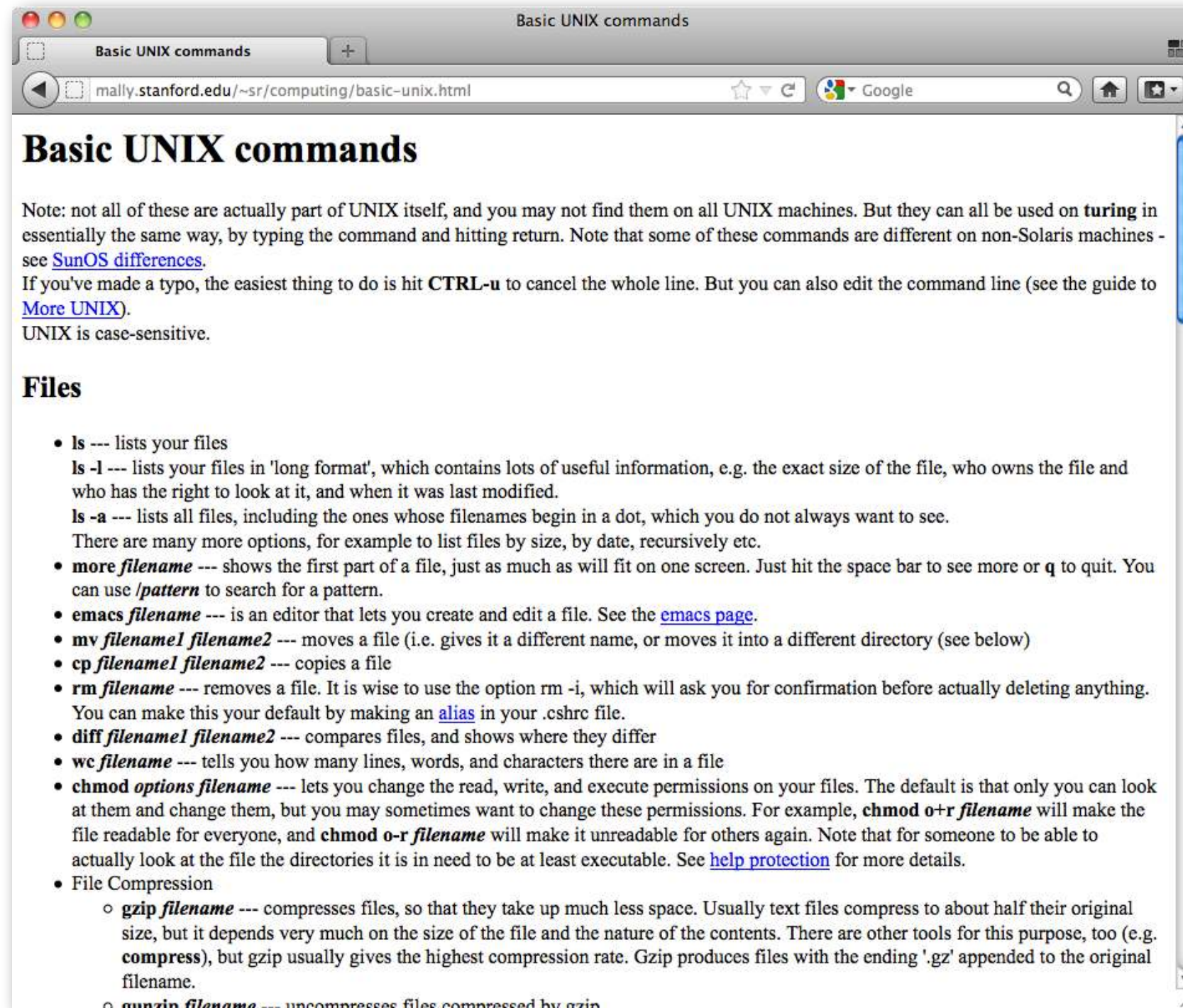
```
ubuntu@ip-10-4-230-31: ~/working — ssh — 141x44
ubuntu@ip-10-4-230-31:~/working$ ls -la ~/
total 444
drwxr-xr-x 18 ubuntu ubuntu 4096 2012-01-09 22:50 .
drwxr-xr-x  6 root    root   4096 2011-11-14 17:12 ..
-rw-r----- 1 ubuntu ubuntu 3757 2012-03-12 12:11 .bash_history
-rw-r--r--  1 ubuntu ubuntu  220 2011-05-18 10:00 .bash_logout
-rw-r--r--  1 ubuntu ubuntu 3581 2011-11-14 17:16 .bashrc
lrwxrwxrwx  1 root    root    21 2011-11-14 12:25 bin -> ../../usr/proftpd/bin
drwxrwxr-x  2 ubuntu ubuntu 4096 2011-11-14 17:16 .byobu
drwxrwxr-x  4 ubuntu ubuntu 4096 2011-11-14 14:33 .cabal
drwx----- 3 ubuntu ubuntu 4096 2011-11-14 11:37 .cache
lrwxrwxrwx  1 root    root    20 2011-11-14 12:23 conf -> ../../usr/nginx/conf
-rwxrwxrwx  1 ubuntu ubuntu  992 2011-11-14 17:12 configure_freenx.sh
drwx----- 3 ubuntu ubuntu 4096 2012-01-08 22:52 .emacs.d
lrwxrwxrwx  1 root    root    21 2011-11-14 12:25 etc -> ../../usr/proftpd/etc
drwxr-xr-x  2 ubuntu ubuntu 4096 2011-11-14 12:51 .fontconfig
drwx----- 2 ubuntu ubuntu 4096 2011-11-14 14:18 .gconf
drwxr-xr-x  3 root    root   4096 2011-11-14 16:58 .gem
drwx----- 2 ubuntu ubuntu 4096 2011-11-14 14:51 .gnupg
lrwxrwxrwx  1 root    root    20 2011-11-14 12:23 html -> ../../usr/nginx/html
lrwxrwxrwx  1 root    root    25 2011-11-14 12:25 include -> ../../usr/proftpd/include
drwxrwxr-x  4 ubuntu ubuntu 4096 2011-11-28 17:49 install
drwxrwxr-x  3 ubuntu ubuntu 4096 2011-11-14 12:27 .lein
-rw-r----- 1 ubuntu ubuntu  65 2011-11-14 13:07 .lessht
lrwxrwxrwx  1 root    root    21 2011-11-14 12:25 lib -> ../../usr/proftpd/lib
lrwxrwxrwx  1 root    root    25 2011-11-14 12:25 libexec -> ../../usr/proftpd/libexec
lrwxrwxrwx  1 root    root    20 2011-11-14 12:23 logs -> ../../usr/nginx/logs
drwxrwxr-x  2 ubuntu ubuntu 4096 2011-11-14 12:51 .m2
drwxrwxr-x  2 ubuntu ubuntu 4096 2011-11-14 14:18 .matplotlib
-rw-r----- 1 ubuntu ubuntu 2964 2012-01-09 22:50 .mysql_history
-rw-r--r--  1 ubuntu ubuntu  675 2011-11-14 17:16 .profile
drwxr-xr-x  2 root    root   4096 2011-11-14 12:25 sbin
-rw-rw-r--  1 ubuntu ubuntu  0 2011-11-14 11:37 .screenrc
lrwxrwxrwx  1 root    root    23 2011-11-14 12:25 share -> ../../usr/proftpd/share
drwx----- 2 ubuntu ubuntu 4096 2011-11-14 11:33 .ssh
-rw-rw-r--  1 ubuntu ubuntu 338416 2012-01-09 16:12 stacks-0.998.tar.gz
drwxrwxr-x  3 ubuntu ubuntu 4096 2011-11-14 13:44 .subversion
-rw-r--r--  1 ubuntu ubuntu  0 2011-11-14 11:38 .sudo_as_admin_successful
drwxrwxr-x  2 ubuntu ubuntu 4096 2012-01-09 04:42 tmp
lrwxrwxrwx  1 root    root    21 2011-11-14 12:25 var -> ../../usr/proftpd/var
-rw-r----- 1 ubuntu ubuntu 7522 2012-01-09 20:35 .viminfo
lrwxrwxrwx  1 ubuntu ubuntu  12 2012-01-08 22:49 working -> /mnt/working
-rw-r----- 1 ubuntu ubuntu  196 2011-12-12 12:04 .Xauthority
ubuntu@ip-10-4-230-31:~/working$
```

Make it comfortable to work in:

- Resize the window
- Change your font size
- Open multiple terminal windows

Obtain a cheat sheet

google “unix commands”



The screenshot shows a web browser window with the title 'Basic UNIX commands'. The address bar displays 'mally.stanford.edu/~sr/computing/basic-unix.html'. The page content is titled 'Basic UNIX commands' and includes a note about command availability, instructions on using CTRL-u to cancel a line, and a section on file commands. The file commands section lists: ls, ls -l, ls -a, more filename, emacs filename, mv filename1 filename2, cp filename1 filename2, rm filename, diff filename1 filename2, wc filename, and chmod options filename. It also includes a subsection for File Compression with gzip and gunzip commands.

Basic UNIX commands

Note: not all of these are actually part of UNIX itself, and you may not find them on all UNIX machines. But they can all be used on **turing** in essentially the same way, by typing the command and hitting return. Note that some of these commands are different on non-Solaris machines - see [SunOS differences](#).
If you've made a typo, the easiest thing to do is hit **CTRL-u** to cancel the whole line. But you can also edit the command line (see the guide to [More UNIX](#)).
UNIX is case-sensitive.

Files

- **ls** --- lists your files
 - ls -l** --- lists your files in 'long format', which contains lots of useful information, e.g. the exact size of the file, who owns the file and who has the right to look at it, and when it was last modified.
 - ls -a** --- lists all files, including the ones whose filenames begin in a dot, which you do not always want to see.
- There are many more options, for example to list files by size, by date, recursively etc.
- **more filename** --- shows the first part of a file, just as much as will fit on one screen. Just hit the space bar to see more or **q** to quit. You can use */pattern* to search for a pattern.
- **emacs filename** --- is an editor that lets you create and edit a file. See the [emacs page](#).
- **mv filename1 filename2** --- moves a file (i.e. gives it a different name, or moves it into a different directory (see below))
- **cp filename1 filename2** --- copies a file
- **rm filename** --- removes a file. It is wise to use the option **rm -i**, which will ask you for confirmation before actually deleting anything. You can make this your default by making an [alias](#) in your `.cshrc` file.
- **diff filename1 filename2** --- compares files, and shows where they differ
- **wc filename** --- tells you how many lines, words, and characters there are in a file
- **chmod options filename** --- lets you change the read, write, and execute permissions on your files. The default is that only you can look at them and change them, but you may sometimes want to change these permissions. For example, **chmod o+r filename** will make the file readable for everyone, and **chmod o-r filename** will make it unreadable for others again. Note that for someone to be able to actually look at the file the directories it is in need to be at least executable. See [help protection](#) for more details.
- File Compression
 - **gzip filename** --- compresses files, so that they take up much less space. Usually text files compress to about half their original size, but it depends very much on the size of the file and the nature of the contents. There are other tools for this purpose, too (e.g. **compress**), but gzip usually gives the highest compression rate. Gzip produces files with the ending '.gz' appended to the original filename.
 - **gunzip filename** --- uncompresses files compressed by gzip

In UNIX everything is a file organized in a hierarchy

Paths

`/home/catchen/working`

Create a series of directories

```
ubuntu@ip-10-4-193-188:~$ mkdir shell
ubuntu@ip-10-4-193-188:~$ cd shell
ubuntu@ip-10-4-193-188:~/shell$ mkdir research
ubuntu@ip-10-4-193-188:~/shell$ ls
research
ubuntu@ip-10-4-193-188:~/shell$ cd research/
ubuntu@ip-10-4-193-188:~/shell/research$ mkdir seq
ubuntu@ip-10-4-193-188:~/shell/research$ ls
seq
ubuntu@ip-10-4-193-188:~/shell/research$ cd seq/
ubuntu@ip-10-4-193-188:~/shell/research/seq$ mkdir radtags
ubuntu@ip-10-4-193-188:~/shell/research/seq$ cd radtags/
ubuntu@ip-10-4-193-188:~/shell/research/seq/radtags$ ls
ubuntu@ip-10-4-193-188:~/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-10-4-193-188:~/shell/research/seq/radtags$ pwd
/home/ubuntu/shell/research/seq/radtags
ubuntu@ip-10-4-193-188:~/shell/research/seq/radtags$
```

% mkdir shell

% cd shell

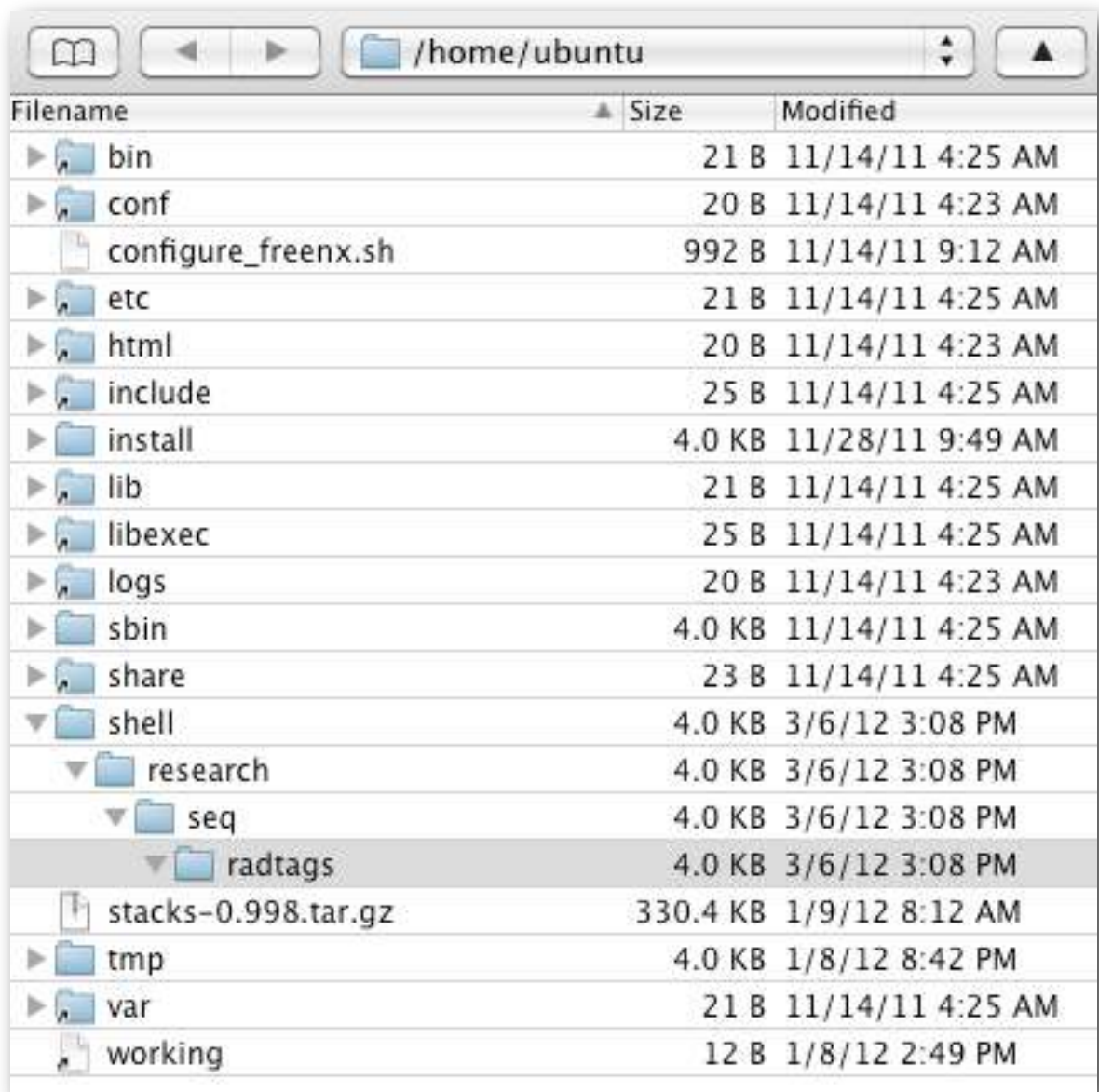
% ls

Paths, cont

This shell view of the nested directories shell, research, seq, and radtags.....

```
ubuntu@ip-10-4-193-188:~$ mkdir shell
ubuntu@ip-10-4-193-188:~$ cd shell
ubuntu@ip-10-4-193-188:~/shell$ mkdir research
ubuntu@ip-10-4-193-188:~/shell$ ls
research
ubuntu@ip-10-4-193-188:~/shell$ cd research/
ubuntu@ip-10-4-193-188:~/shell/research$ mkdir seq
ubuntu@ip-10-4-193-188:~/shell/research$ ls
seq
ubuntu@ip-10-4-193-188:~/shell/research$ cd seq/
ubuntu@ip-10-4-193-188:~/shell/research/seq$ mkdir radtags
ubuntu@ip-10-4-193-188:~/shell/research/seq$ cd radtags/
ubuntu@ip-10-4-193-188:~/shell/research/seq/radtags$ ls
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-10-4-193-188:~/shell/research/seq/radtags$ pwd
/home/ubuntu/shell/research/seq/radtags
ubuntu@ip-10-4-193-188:~/shell/research/seq/radtags$
```

.... is equivalent to this GUI view of the same directories

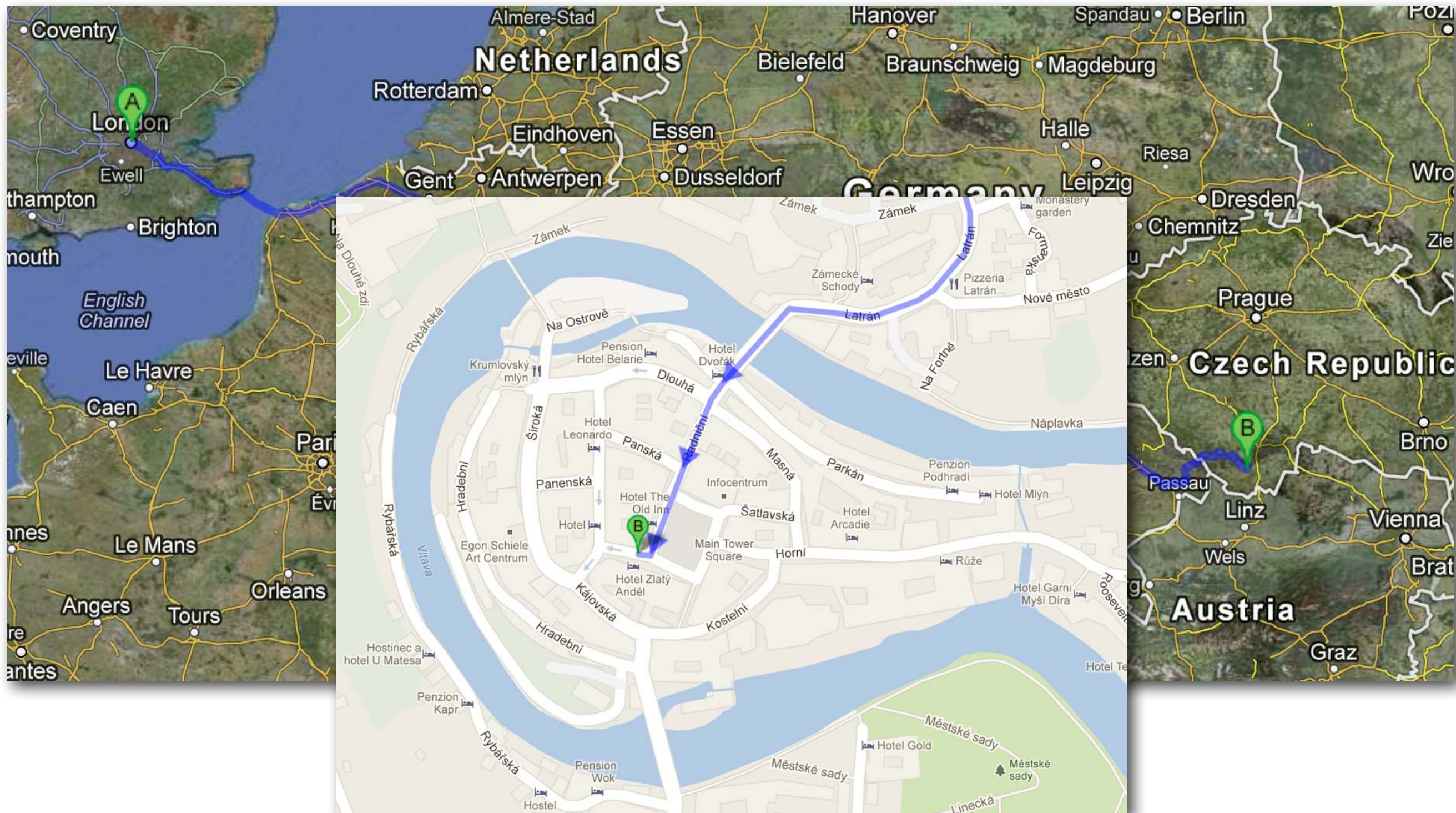


Filename	Size	Modified
bin	21 B	11/14/11 4:25 AM
conf	20 B	11/14/11 4:23 AM
configure_freenx.sh	992 B	11/14/11 9:12 AM
etc	21 B	11/14/11 4:25 AM
html	20 B	11/14/11 4:23 AM
include	25 B	11/14/11 4:25 AM
install	4.0 KB	11/28/11 9:49 AM
lib	21 B	11/14/11 4:25 AM
libexec	25 B	11/14/11 4:25 AM
logs	20 B	11/14/11 4:23 AM
sbin	4.0 KB	11/14/11 4:25 AM
share	23 B	11/14/11 4:25 AM
shell	4.0 KB	3/6/12 3:08 PM
research	4.0 KB	3/6/12 3:08 PM
seq	4.0 KB	3/6/12 3:08 PM
radtags	4.0 KB	3/6/12 3:08 PM
stacks-0.998.tar.gz	330.4 KB	1/9/12 8:12 AM
tmp	4.0 KB	1/8/12 8:42 PM
var	21 B	11/14/11 4:25 AM
working	12 B	1/8/12 2:49 PM

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

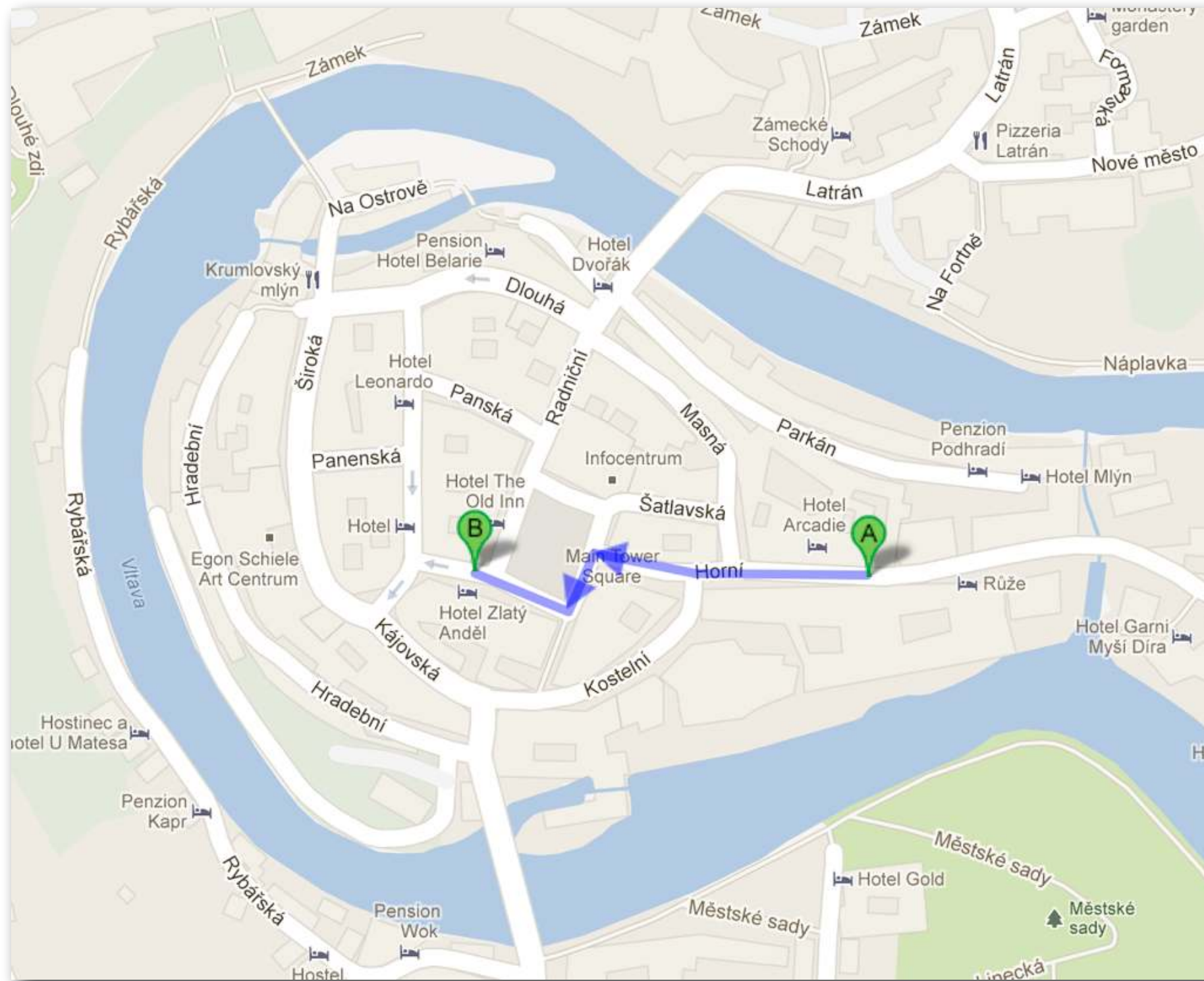
Absolute and relative paths

How do I get to the Hotel Zlatý Anděl?

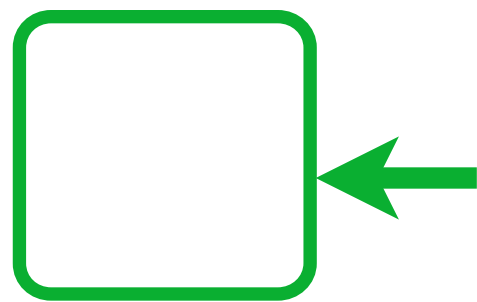


Absolute and relative paths

How do I get to the Hotel Zlatý Andel?



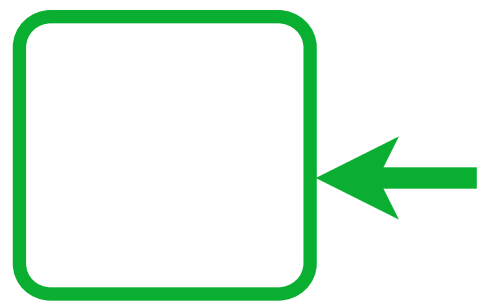
Paths



Absolute Path

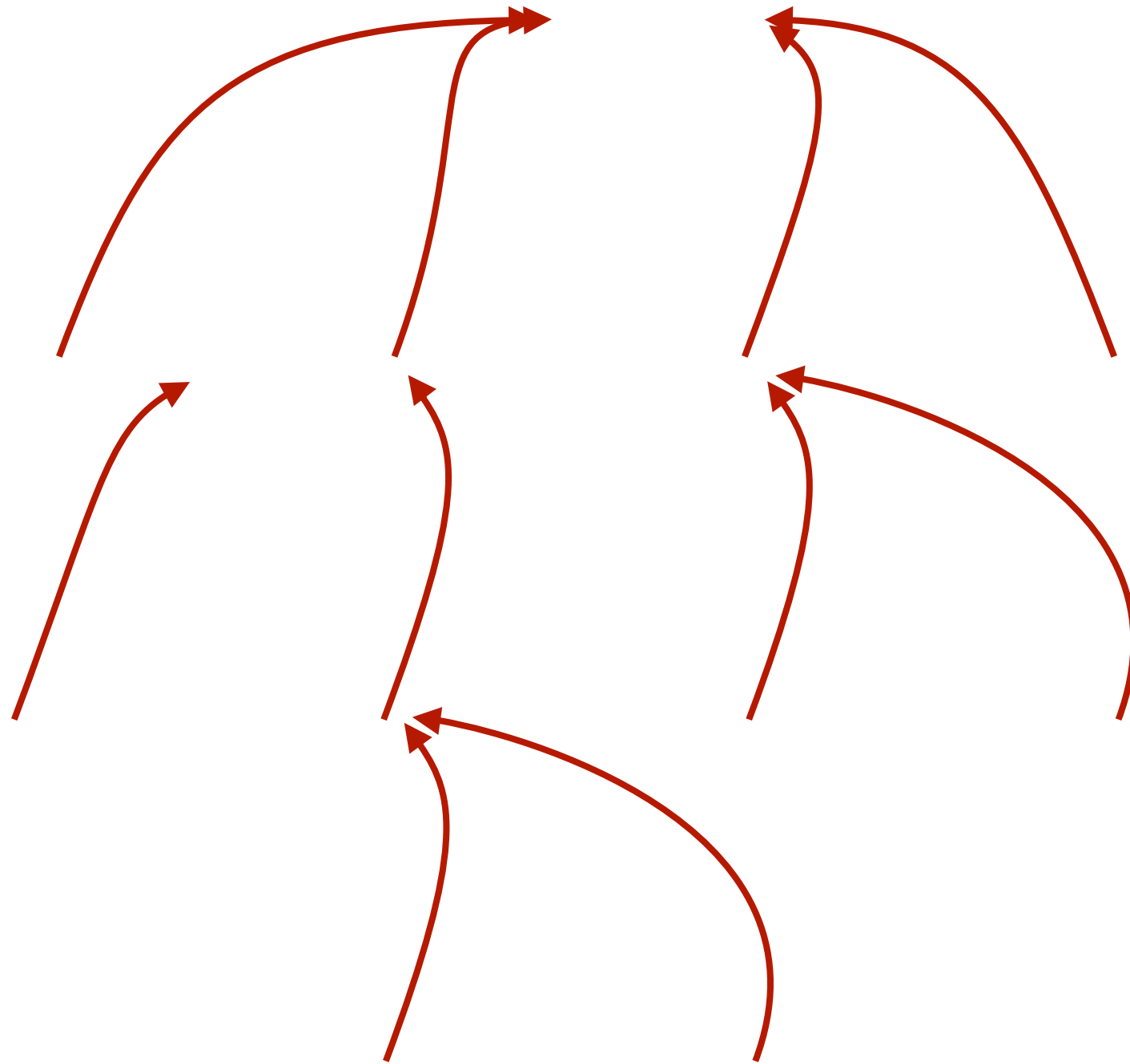
`/home/catchen/working`

Relative Path?

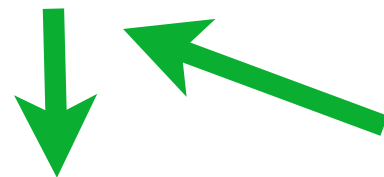


Special files -- '*dot*'

Special files -- '*dot dot*'



Relative Path



`../working`

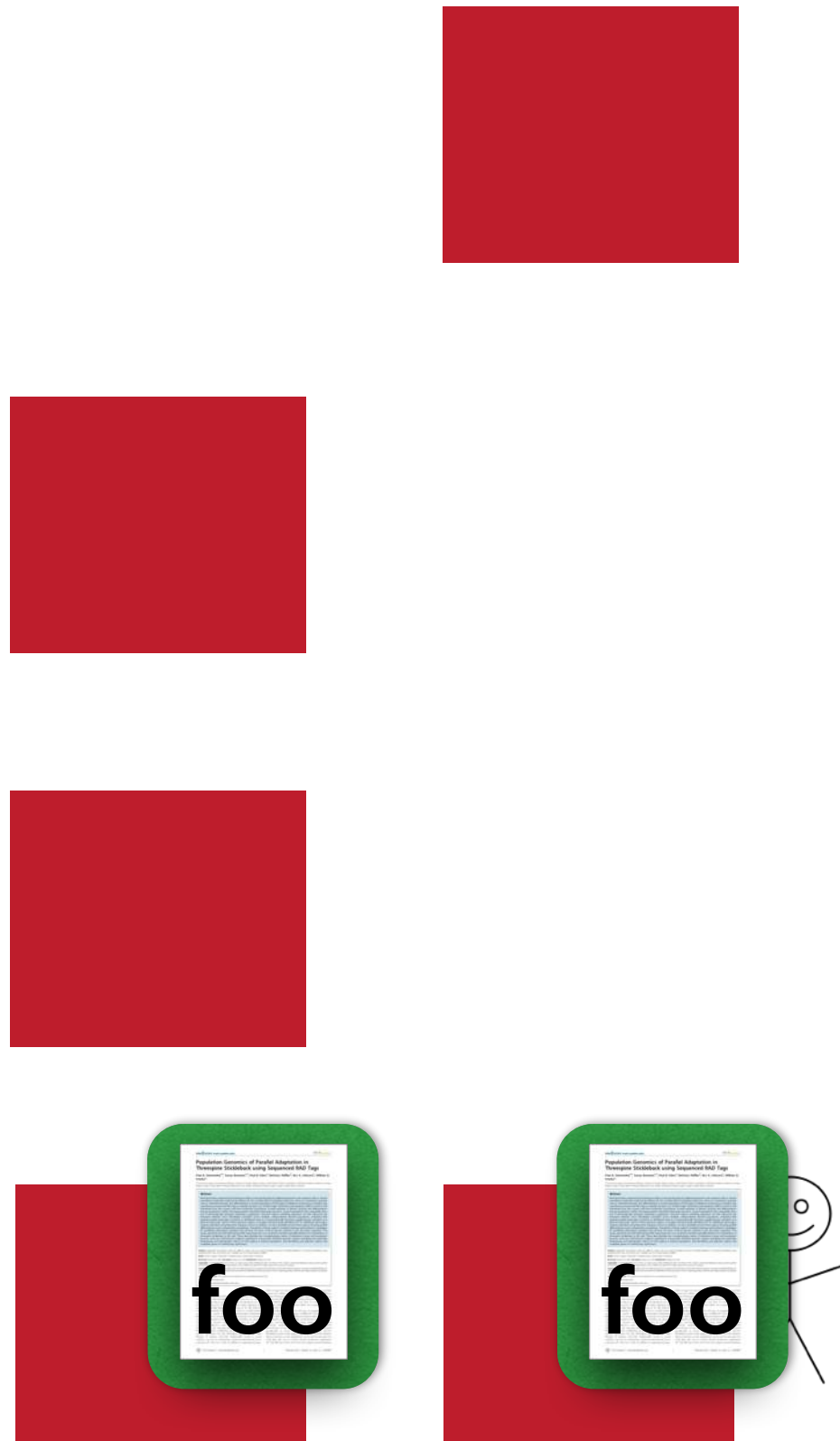


Absolute Path:

/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

```
ubuntu@ip-10-4-193-188:~$ mkdir shell
ubuntu@ip-10-4-193-188:~$ cd shell
ubuntu@ip-10-4-193-188:~/shell$ mkdir research
ubuntu@ip-10-4-193-188:~/shell$ ls
research
ubuntu@ip-10-4-193-188:~/shell$ cd research/
ubuntu@ip-10-4-193-188:~/shell/research$ mkdir seq
ubuntu@ip-10-4-193-188:~/shell/research$ ls
seq
ubuntu@ip-10-4-193-188:~/shell/research$ cd seq/
ubuntu@ip-10-4-193-188:~/shell/research/seq$ mkdir radtags
ubuntu@ip-10-4-193-188:~/shell/research/seq$ cd radtags/
ubuntu@ip-10-4-193-188:~/shell/research/seq/radtags$ ls
ubuntu@ip-10-4-193-188:~/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-10-4-193-188:~/shell/research/seq/radtags$ pwd
/home/ubuntu/shell/research/seq/radtags
ubuntu@ip-10-4-193-188:~/shell/research/seq/radtags$
```

Special Files

dot

dot dot

% ls .

% ls ..

% ls ../../

Binary programs - ls, cp, mkdir, etc.

```
ubuntu@ip-10-140-6-74: ~ — ssh — 133x24
ubuntu@ip-10-140-6-74:~$ ls /bin
bash      csh      getfacl  lsblk    ntfsck   readlink  true
bunzip2   dash     grep     lsmod    ntfscluster  rm      ulockmgr_server
busybox   date     gunzip   mkdir    ntfscomp  rmdir    umount
bzipcat   dbus-cleanup-sockets gzexe    mknod    ntfsdecrypt  nano    uname
bzip      dbus-daemon  gzip     mktemp   ntfsdump_logfile  run-parts  uncompress
bzdiff    dbus-uuidgen  hostname  more      ntfsfix     rzsh     unicode_start
bzegrep   dd        init-checkconf  mount      ntfsinfo    sed      vdir
bzexe     df         initctl2dot  mountpoint  ntfsls      setfacl  which
bzfgrep   dir       ip          mt         ntfsmftalloc  setfont  ypdomainname
bzgrep    dmesg     kbd_mode   mt-gnu     ntfsmove     sh       zcat
bzip2     dnsdomainname kill        mv         ntfstruncate  sh.distrib  zcmp
bzip2recover domainname  ksh        nano       ntfswipe     sleep    zdiff
bzless    dumpkeys  less       nc         open          static-sh  zegrep
bzip      echo      lessecho   nc.openbsd  openvt       stty     zfgrep
bzipmore  ed        lessfile   netcat     pidof        su       zforce
cat        egrep     lesskey    netstat    ping         sync     zgrep
chacl     false    lesspipe   nisdomainname  ping6      tailf    zless
chgrp     fgconsole ln          ntfs-3g    plymouth    tar      zmore
chmod     fgrep    loadkeys   ntfs-3g.probe  plymouth-upstart-bridge  tempfile  znew
chvt      findmnt  login      ntfs-3g.secaudit  ps         tcsh     zsh
cp        fuser    lowntfs-3g ntfs-3g.usermap  pwd        touch    zsh4
cpio      fusermount ls          ntfsckat    rbash
```

% 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

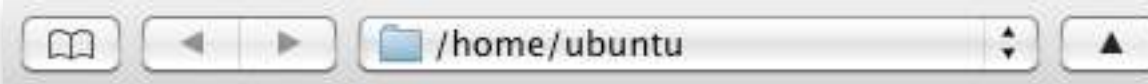
% pwd

Relative and absolute paths

/home/ubuntu/shell/research/seq/radtags

Create a series of directories under shell:

```
ubuntu@ip-10-4-193-188:~$ mkdir shell
ubuntu@ip-10-4-193-188:~$ cd shell
ubuntu@ip-10-4-193-188:~/shell$ mkdir research
ubuntu@ip-10-4-193-188:~/shell$ ls
research
ubuntu@ip-10-4-193-188:~/shell$ cd research/
ubuntu@ip-10-4-193-188:~/shell/research$ mkdir seq
ubuntu@ip-10-4-193-188:~/shell/research$ ls
seq
ubuntu@ip-10-4-193-188:~/shell/research$ cd seq/
ubuntu@ip-10-4-193-188:~/shell/research/seq$ mkdir radtags
ubuntu@ip-10-4-193-188:~/shell/research/seq$ cd radtags/
ubuntu@ip-10-4-193-188:~/shell/research/seq/radtags$ ls
ubuntu@ip-10-4-193-188:~/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-10-4-193-188:~/shell/research/seq/radtags$ pwd
/home/ubuntu/shell/research/seq/radtags
ubuntu@ip-10-4-193-188:~/shell/research/seq/radtags$
```



Filename	Size	Modified
bin	21 B	11/14/11 4:25 AM
conf	20 B	11/14/11 4:23 AM
configure_freenx.sh	992 B	11/14/11 9:12 AM
etc	21 B	11/14/11 4:25 AM
html	20 B	11/14/11 4:23 AM
include	25 B	11/14/11 4:25 AM
install	4.0 KB	11/28/11 9:49 AM
lib	21 B	11/14/11 4:25 AM
libexec	25 B	11/14/11 4:25 AM
logs	20 B	11/14/11 4:23 AM
sbin	4.0 KB	11/14/11 4:25 AM
share	23 B	11/14/11 4:25 AM
shell	4.0 KB	3/6/12 3:08 PM
research	4.0 KB	3/6/12 3:08 PM
seq	4.0 KB	3/6/12 3:08 PM
radtags	4.0 KB	3/6/12 3:08 PM
stacks-0.998.tar.gz	330.4 KB	1/9/12 8:12 AM
tmp	4.0 KB	1/8/12 8:42 PM
var	21 B	11/14/11 4:25 AM
working	12 B	1/8/12 2:49 PM

% 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>
```

```
root@ubuntu:~# cd /etc/
root@ubuntu:/etc# ls -la
total 788
drwxr-xr-x 82 root root 4096 2012-06-25 01:56 .
drwxr-xr-x 23 root root 4096 2012-04-18 09:17 ..
-rw-r--r-- 1 root root 2981 2012-04-16 10:09 adduser.conf
drwxr-xr-x 2 root root 4096 2012-04-16 10:23 alternatives
drwxr-xr-x 3 root root 4096 2012-04-16 10:19 apt
drwxr-xr-x 3 root root 4096 2012-04-16 10:21 apparmor
drwxr-xr-x 8 root root 4096 2012-04-16 12:09 apparmor.d
drwxr-xr-x 3 root root 4096 2012-04-16 12:09 apport
drwxr-xr-x 6 root root 4096 2012-04-16 12:08 apt
-rw-r----- 1 root daemon 144 2011-05-16 03:32 at.deny
-rw-r--r-- 1 root root 1975 2011-05-18 03:00 bash.bashrc
-rw-r--r-- 1 root root 58753 2011-10-04 13:53 bash_completion
drwxr-xr-x 3 root root 12288 2012-04-16 12:09 bash_completion.d
-rw-r--r-- 1 root root 344 2011-10-04 12:56 bindresvport.blacklist
-rw-r--r-- 1 root root 321 2011-08-09 09:16 blkid.conf
lrwxrwxrwx 1 root root 15 2011-08-09 09:16 blkid.tab -> /dev/blkid.tab
drwxr-xr-x 2 root root 4096 2012-04-16 10:23 byobu
drwxr-xr-x 3 root root 4096 2012-04-16 10:18 ca-certificates
-rw-r--r-- 1 root root 7014 2012-04-16 10:21 ca-certificates.conf
drwxr-xr-x 2 root root 4096 2012-04-16 10:22 calendar
drwxr-s--- 2 root dip 4096 2012-04-16 10:22 chatscripts
drwxr-xr-x 4 root root 4096 2012-04-16 12:10 cloud
drwxr-xr-x 2 root root 4096 2012-04-16 10:10 console-setup
drwxr-xr-x 2 root root 4096 2012-04-16 10:10 cron.d
drwxr-xr-x 2 root root 4096 2012-04-16 12:09 cron.daily
drwxr-xr-x 2 root root 4096 2012-04-16 10:10 cron.hourly
drwxr-xr-x 2 root root 4096 2012-04-16 10:10 cron.monthly
-rw-r--r-- 1 root root 724 2011-09-19 17:04 crontab
drwxr-xr-x 2 root root 4096 2012-04-16 10:22 cron.weekly
drwxr-xr-x 4 root root 4096 2012-04-16 10:21 dbus-1
-rw-r--r-- 1 root root 2969 2011-06-23 00:01 debconf.conf
-rw-r--r-- 1 root root 11 2011-07-08 10:13 debian_version
drwxr-xr-x 2 root root 4096 2012-04-16 12:09 default
```

Three variants to ls

```
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

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

Explore the file hierarchy

1. Move to the directory `/etc`

- What is the first line of the files `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:

`/unix_data/batch_1.genotypes_1.loc.gz`

FASTQ file:

`/unix_data/s_1_sequence.txt.gz`

Tar Archive:

`/unix_data/samples.tar.gz`

What is a tar archive?



tar = tape archive

Compress / Decompress

`gzip / gunzip`

`batch_l.genotypes_l.loc.gz`

`s_l_sequence.txt.gz`

Gzipped Tar archive

`tar xvfz`

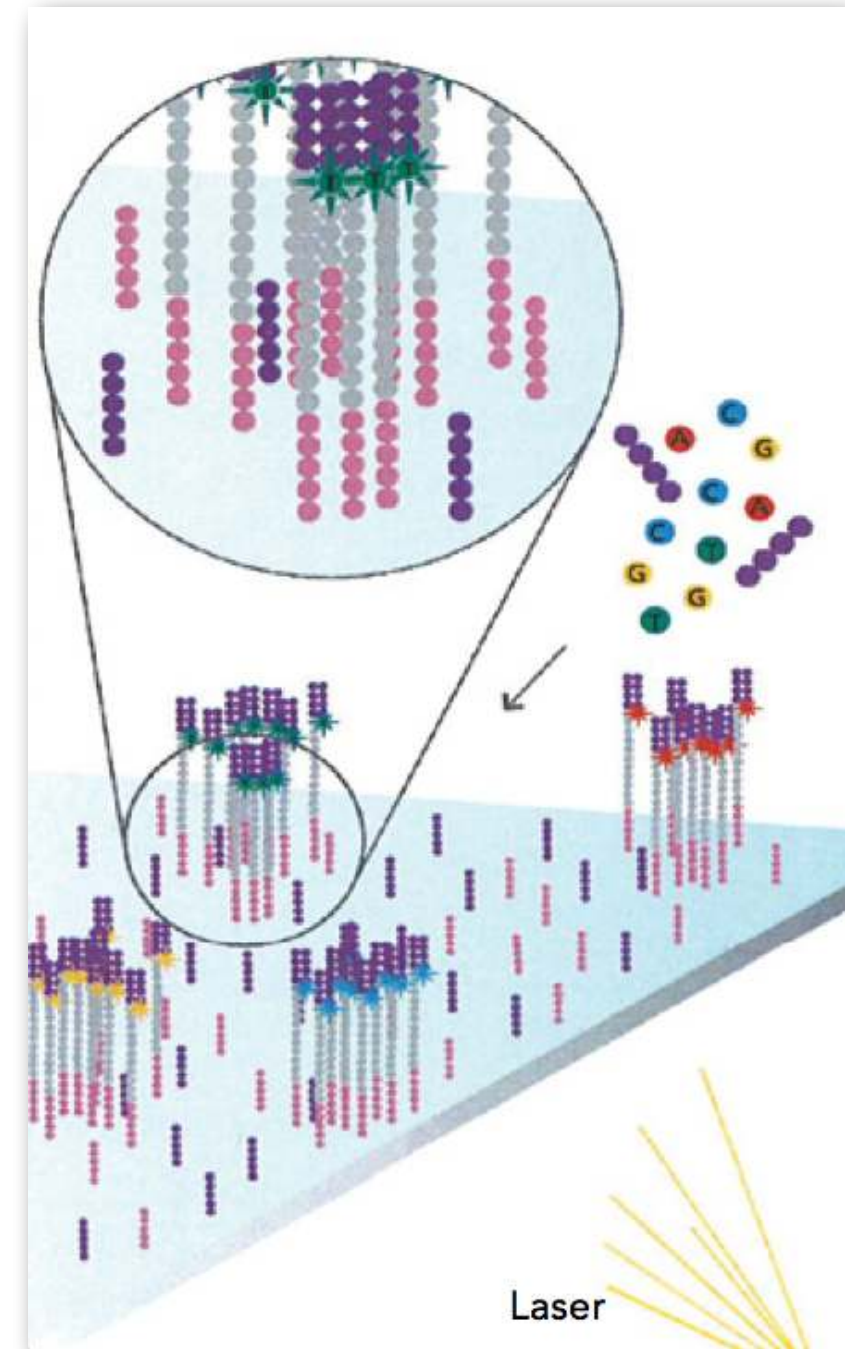
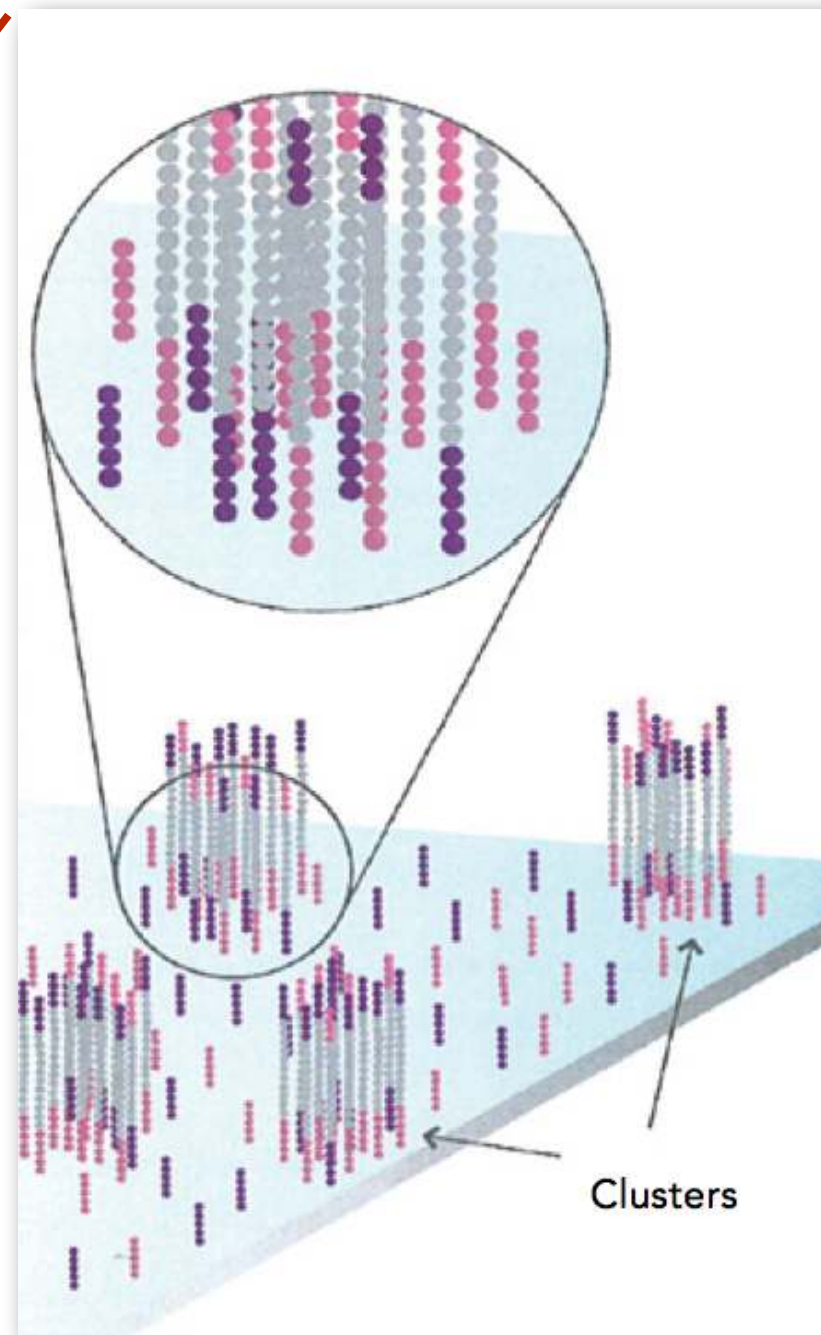
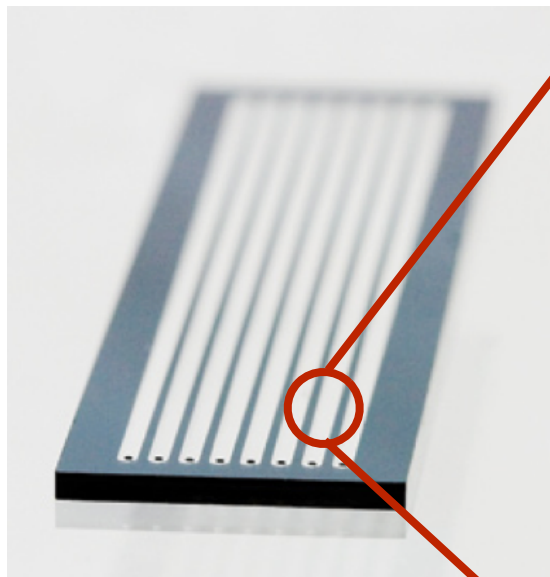
`samples.tar.gz`

Tar archive

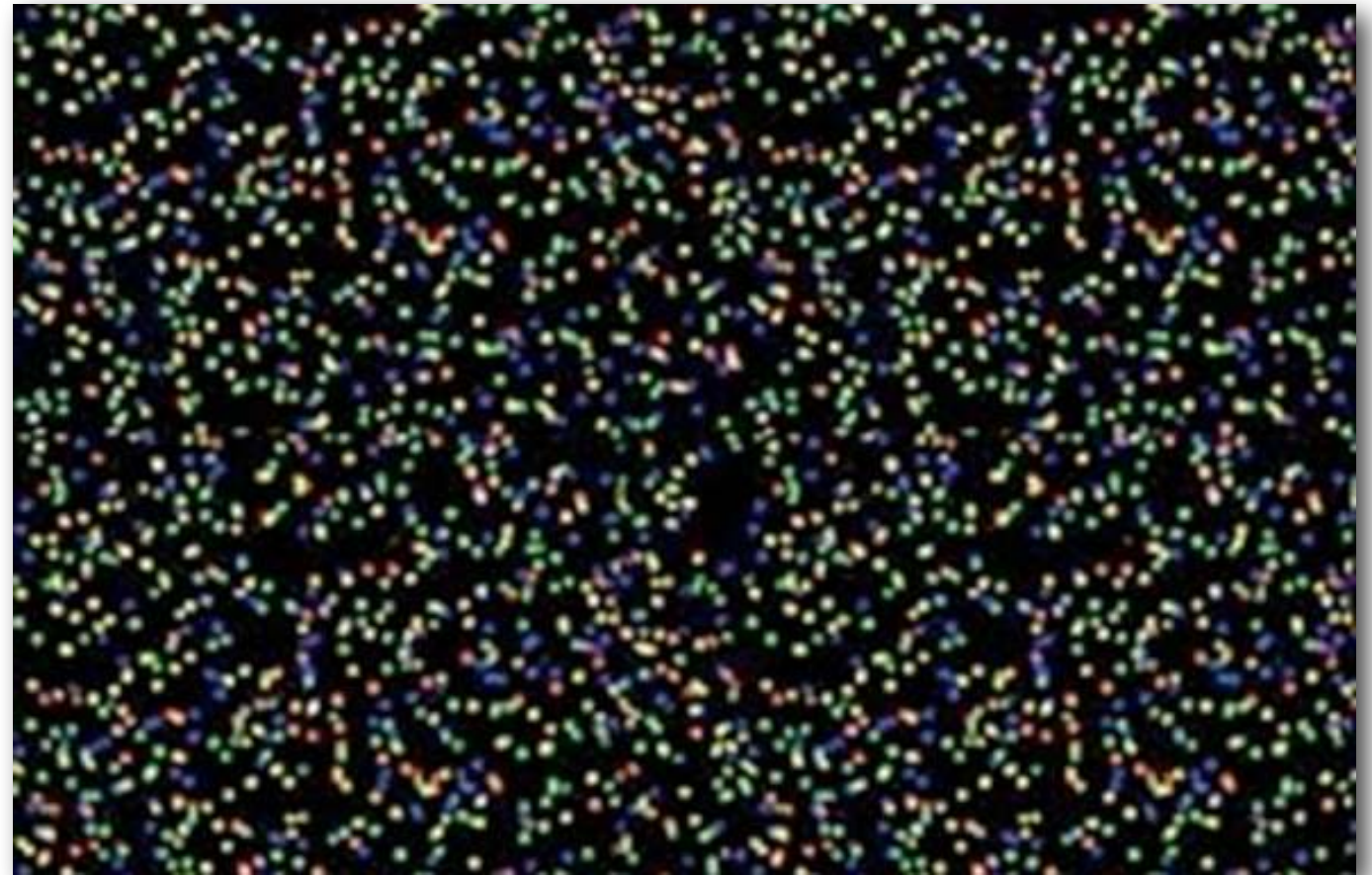
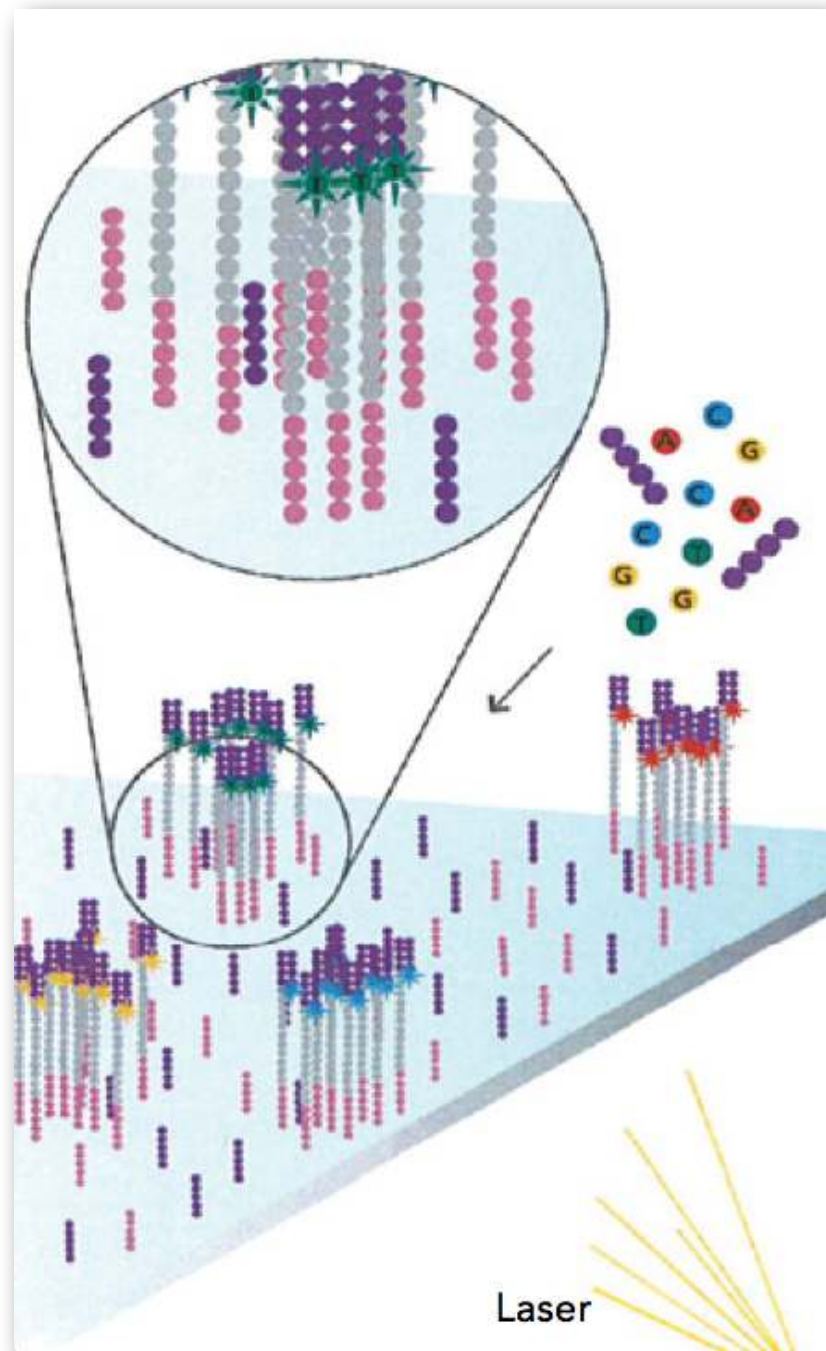
`tar xvf`

`samples.tar`

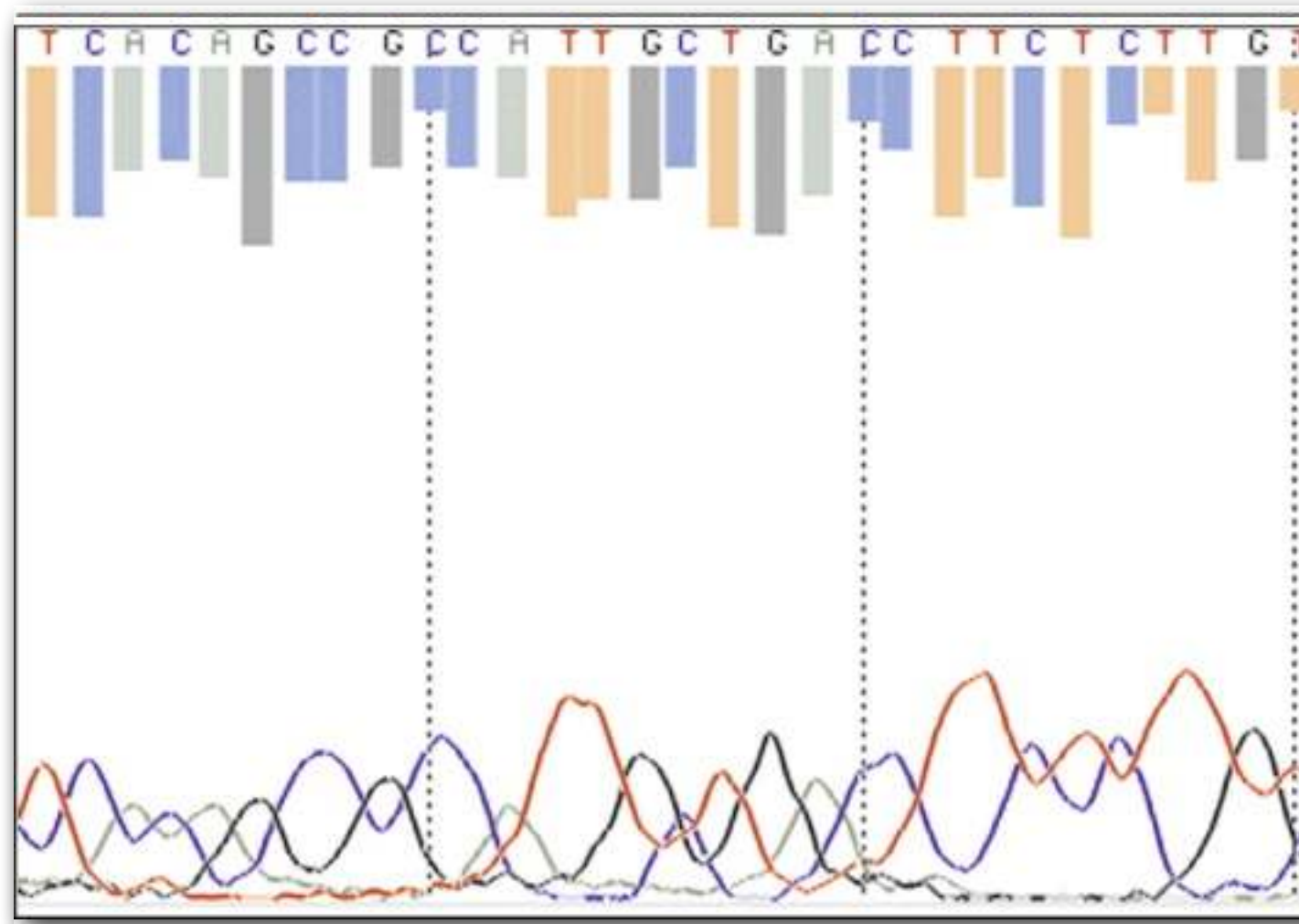
Sequencing on Illumina's Flow cell



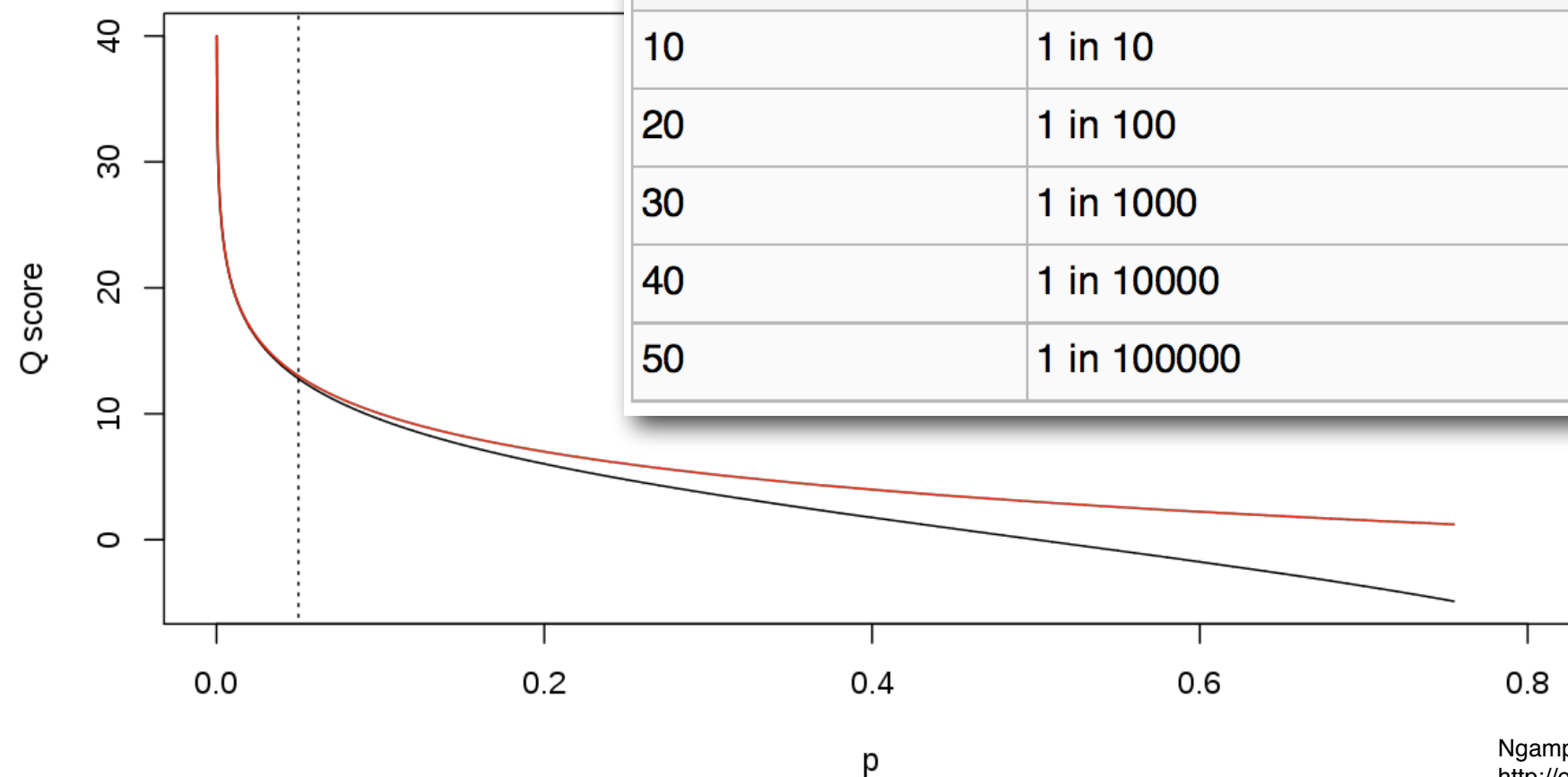
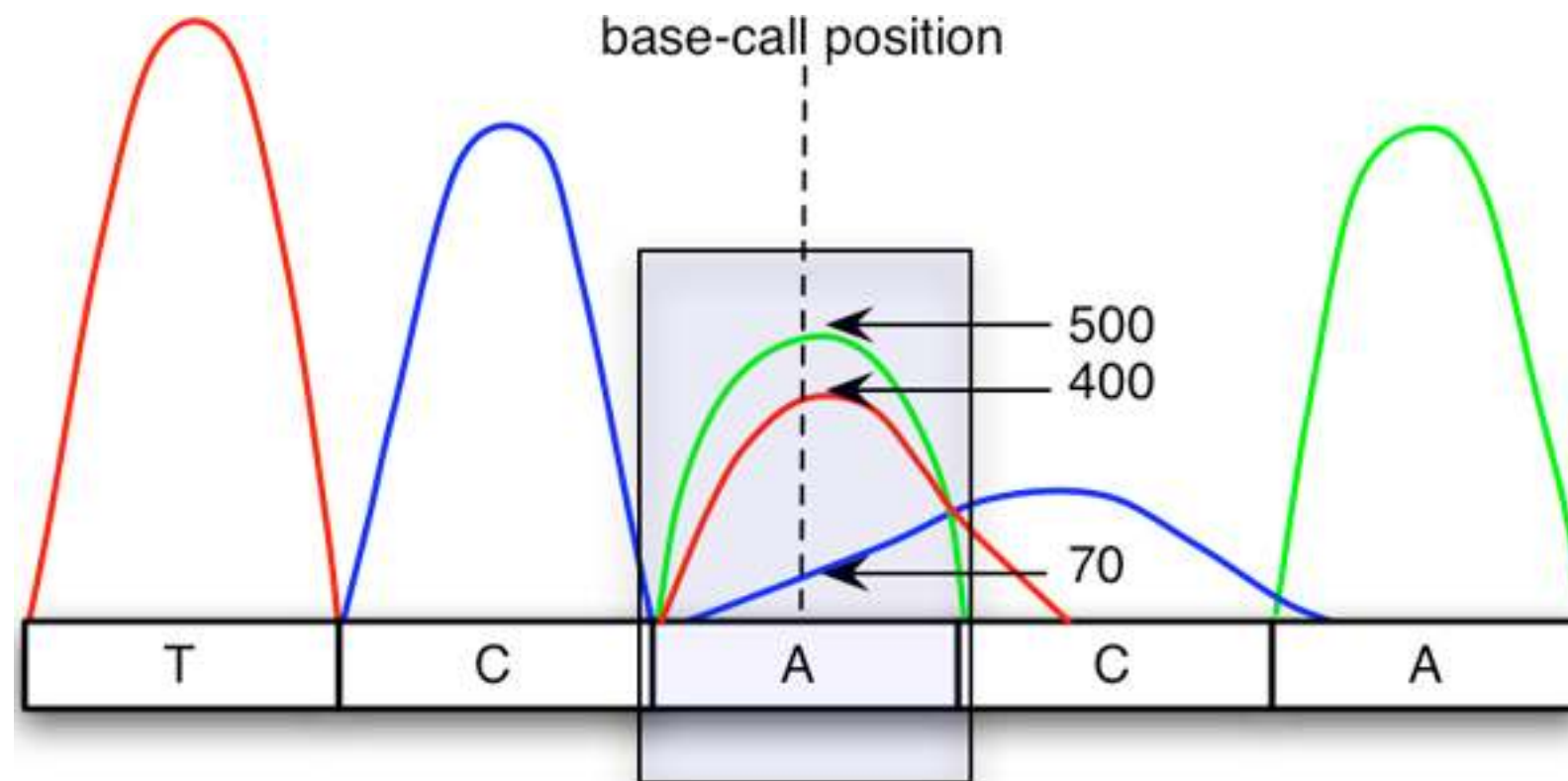
Sequencing on Illumina's Flow cell, ctd.



Sequencing on Illumina's Flow cell, ctd.



Phred Quality Score



Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

The FASTQ File Format

FASTA

```
>chromosome7
TTTGTCTGCAGGGGGACACGTCAAAGTCAAACGCAGGCAAGTTTGTGTTTATGTCCAGTGGATCTTTTGATTTT
ACATACTGCAGGGGTCAGGAGGATTATCTCCTCTGCAAGGTAACGCCTGCTGTAACCGTTGTTCTTCATCCTTTT
CCTAACTGCAGGGGCTGTCTTGTCAGGTCTGACAAGACATATGCAGGGGCTCAATTTGAGATAATTGCTCAATATA
```

FASTQ

```
@Sequence_137
TTTGTCTGCAGGGGGACACGTCAAAGTCAAACGCAGGCAAGTTTGTGTTTATGTCCAGTGGATCTTTTGATTTT
+Sequence_137
<?@DDDDDDHFHHFBB@GGIACFHGGHGBGHGCDHBEAHACHI=@CH.=7ACAHHADECDBCC66(6>@C>5@CACCA
```

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

ASCII Code

0	<NUL>	32	<SPC>	64	@	96	`	128	Ä	160	†	192	¿	224	‡
1	<SOH>	33	!	65	A	97	a	129	Å	161	°	193	ı	225	•
2	<STX>	34	"	66	B	98	b	130	Ç	162	¢	194	¬	226	,
3	<ETX>	35	#	67	C	99	c	131	É	163	£	195	√	227	„
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	Â
6	<ACK>	38	&	70	F	102	f	134	Ü	166	¶	198	Δ	230	Ê
7	<BEL>	39	'	71	G	103	g	135	á	167	ß	199	«	231	Á
8	<BS>	40	(	72	H	104	h	136	à	168	®	200	»	232	Ë
9	<TAB>	41	)	73	I	105	i	137	â	169	©	201	...	233	È
10	<LF>	42	*	74	J	106	j	138	ä	170	™	202		234	Í
11	<VT>	43	+	75	K	107	k	139	ã	171	'	203	À	235	Î
12	<FF>	44	,	76	L	108	l	140	å	172	“	204	Ã	236	Ï
13	<CR>	45	-	77	M	109	m	141	ç	173	≠	205	Ö	237	Ì
14	<SO>	46	.	78	N	110	n	142	é	174	Æ	206	Œ	238	Ó
15	<SI>	47	/	79	O	111	o	143	è	175	Ø	207	œ	239	Ô
16	<DLE>	48	0	80	P	112	p	144	ê	176	∞	208	-	240	Ⓜ
17	<DC1>	49	1	81	Q	113	q	145	ë	177	±	209	—	241	Ò
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	Û
20	<DC4>	52	4	84	T	116	t	148	î	180	¥	212	`	244	Ü
21	<NAK>	53	5	85	U	117	u	149	ï	181	μ	213	'	245	ı
22	<SYN>	54	6	86	V	118	v	150	ñ	182	ð	214	÷	246	ˆ
23	<ETB>	55	7	87	W	119	w	151	ó	183	Σ	215	◇	247	˜
24	<CAN>	56	8	88	X	120	x	152	ò	184	Π	216	ÿ	248	—
25		57	9	89	Y	121	y	153	ô	185	π	217	Ÿ	249	˘
26	<SUB>	58	:	90	Z	122	z	154	ö	186	ƒ	218	/	250	·
27	<ESC>	59	;	91	[	123	{	155	õ	187	ª	219	€	251	°
28	<FS>	60	<	92	\	124		156	ú	188	º	220	<	252	¸
29	<GS>	61	=	93	]	125	}	157	ù	189	Ω	221	>	253	”
30	<RS>	62	>	94	^	126	~	158	û	190	æ	222	fi	254	˚
31	<US>	63	?	95	_	127		159	ü	191	ø	223	fl	255	˛

8 bits = 2^8 combinations = 256

0 1 1 1 1 0 0 1

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

The FASTQ File Format, ctd

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

Quality Scores

[illegible]

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:  
TTTGTCTGCAGGGGGACACGTCAAAGTCAAACGCAGGCAAGTTTGTGTTTATGTCCAGTGGATCTTTTGATTTT  
+  
<?@DDDDDDHFHHFBB@GGIACFHGGHGBGHGCDHBEAHACHI=@CH.=7ACAHHADECDBCC66 ( 6>@C>5@CACCA
```

$$70 - 33 = 37$$

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

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
```

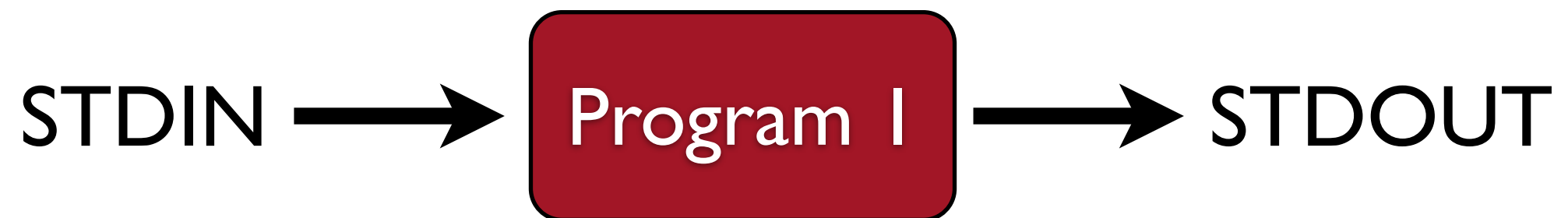
Special Files

STDIN, STDOUT, STDERR

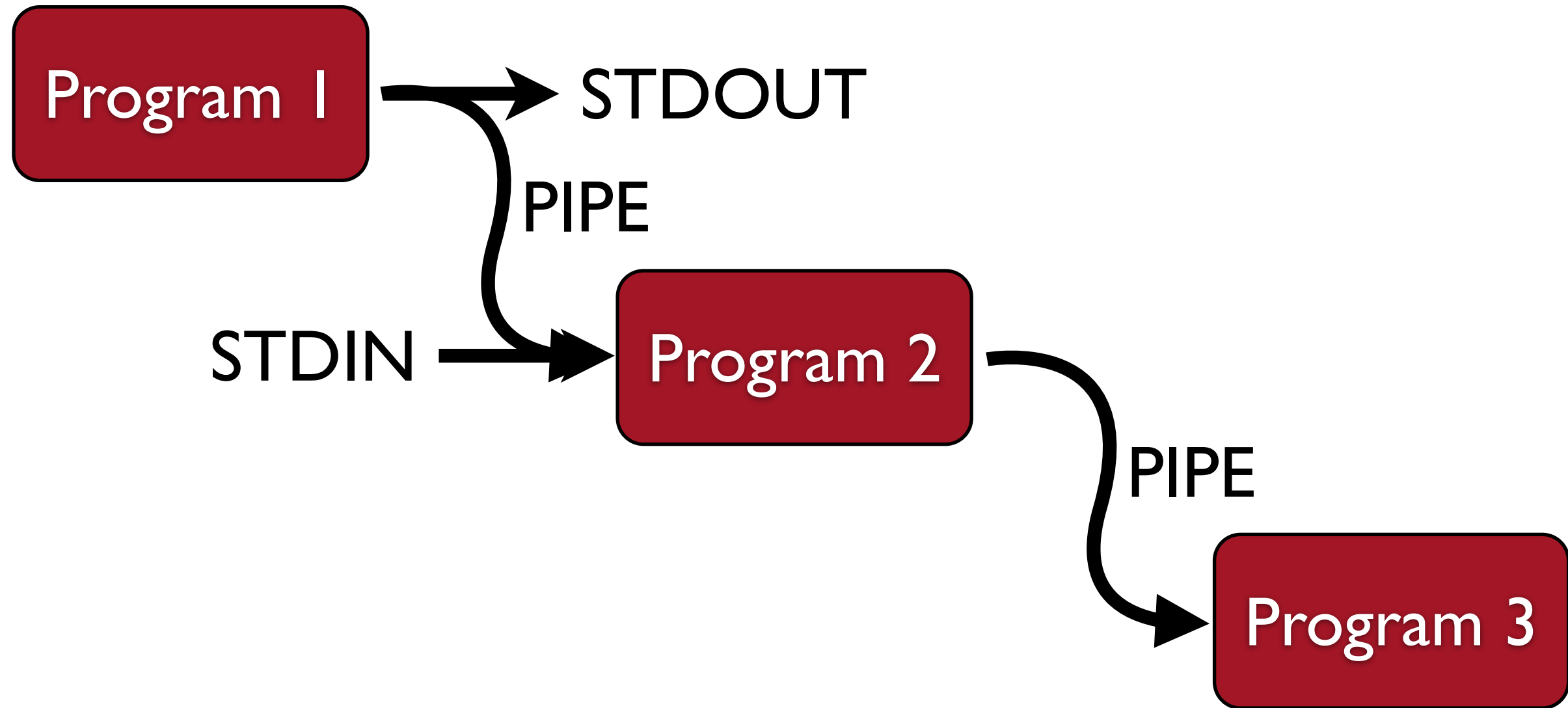
The Shell's Killer App: Pipes



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"
```

s_l_sequence.txt.gz

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

```
ls
gunzip
man
more
cat
wc
head
cut
grep
sort
uniq
>
|
```

```
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

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"
```

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
```

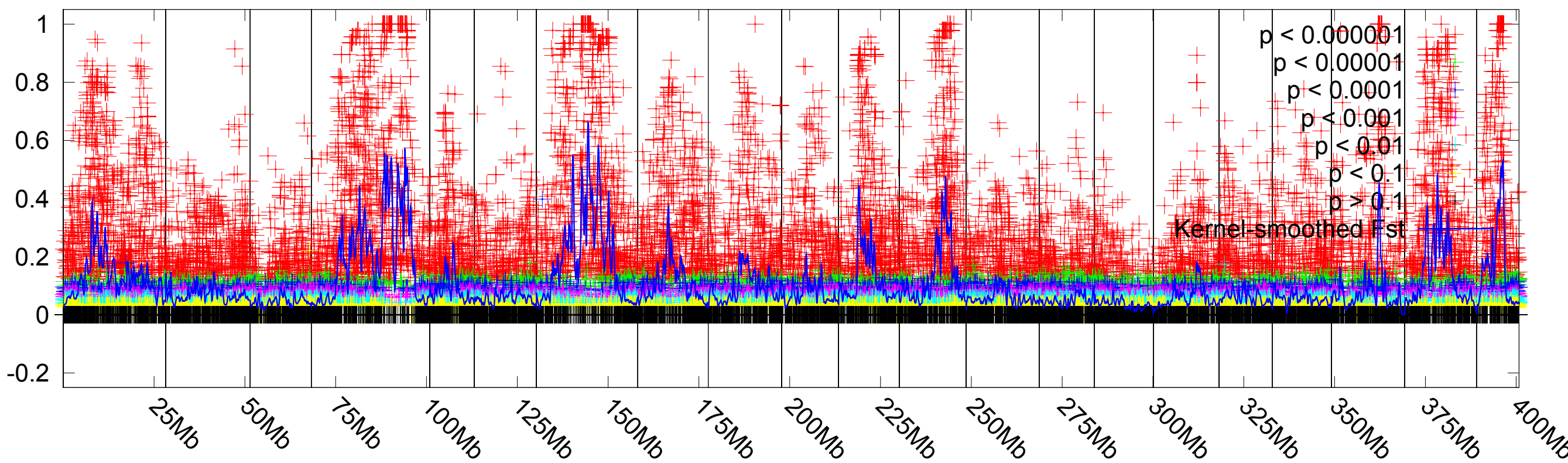
```
grep -v -c "@" s_1_sequence.txt
```

Count reads with barcode:

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

Problem Set #1

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



#	Batch	ID	Locus	ID	Pop 1	ID	Pop 2	ID	Chr	BP	Column	Overall	Pi	Fst	Fisher's	P	Odds	Ratio
2	7894	2	3	groupI	11832	19	0.428182	-0.0076252913						0.191294			0.687192	
2	7896	2	3	groupI	11900	83	0.328622	0.1775694587						4.35747e-08			5.44667	
2	9611	2	3	groupI	49756	48	0.090426	-0.1127451906						0.00109115			0.0898072	
2	9611	2	3	groupI	49765	57	0.0132887	-0.1522407447						0.518395			0.493113	
2	9611	2	3	groupI	49766	58	0.0133776	-0.1563725438						0.520033			0.501401	
...																		
2	7083	2	3	groupIX	20172984	57	0.0480227	-0.1249752728						0.0230205			0.152263	
2	7083	2	3	groupIX	20173004	77	0.0181808	-0.1405892833						0.27288			0.355556	
2	7083	2	3	groupIX	20173016	89	0.0424072	-0.1243625038						0.0442229			0.17037	
2	7083	2	3	groupIX	20173018	91	0.148072	0.2305903127						2.24832e-07			13.2936	
2	7084	2	3	groupIX	20172960	29	0.388239	-0.0102170169						0.307185			0.758001	
2	7084	2	3	groupIX	20172995	64	0.368259	-0.0130602425						0.233889			0.70084	
2	7084	2	3	groupIX	20173013	82	0.0121578	-0.1359057575						0.51309			0.46472	