

Connecting to Your Virtual Ubuntu Server

unix, clients and servers, ssh, scp

Frederick J Tan
Bioinformatics Research Faculty
Carnegie Institution of Washington, Department of Embryology
tan@ciwemb.edu

26 August 2013

Unix, Linux, Ubuntu, Oh My!

UNIX[®]

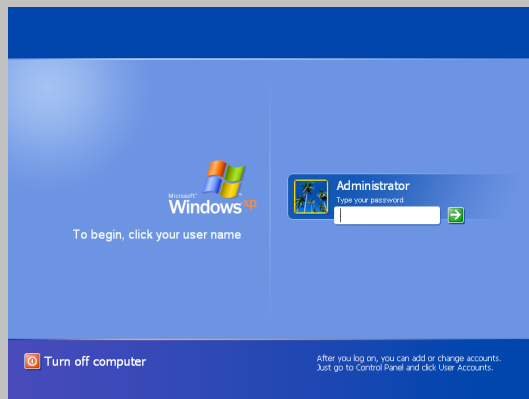


The Awesome Power of the Client-Server Model



How All the Pieces Fit Together

Your Physical Laptop



Virtual Server



```
Ubuntu 12.10 ubuntu tty1
```

```
ubuntu login: _
```

Local Commands



```
smith@laptop$ scp  
fedora:file.txt Desktop
```

Remote Control



```
smith@laptop$ ssh ubuntu
```



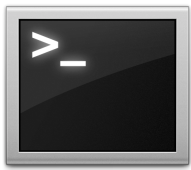
```
smith@ubuntu$ []
```

Department Server



```
Fedora release 9 tty1
```

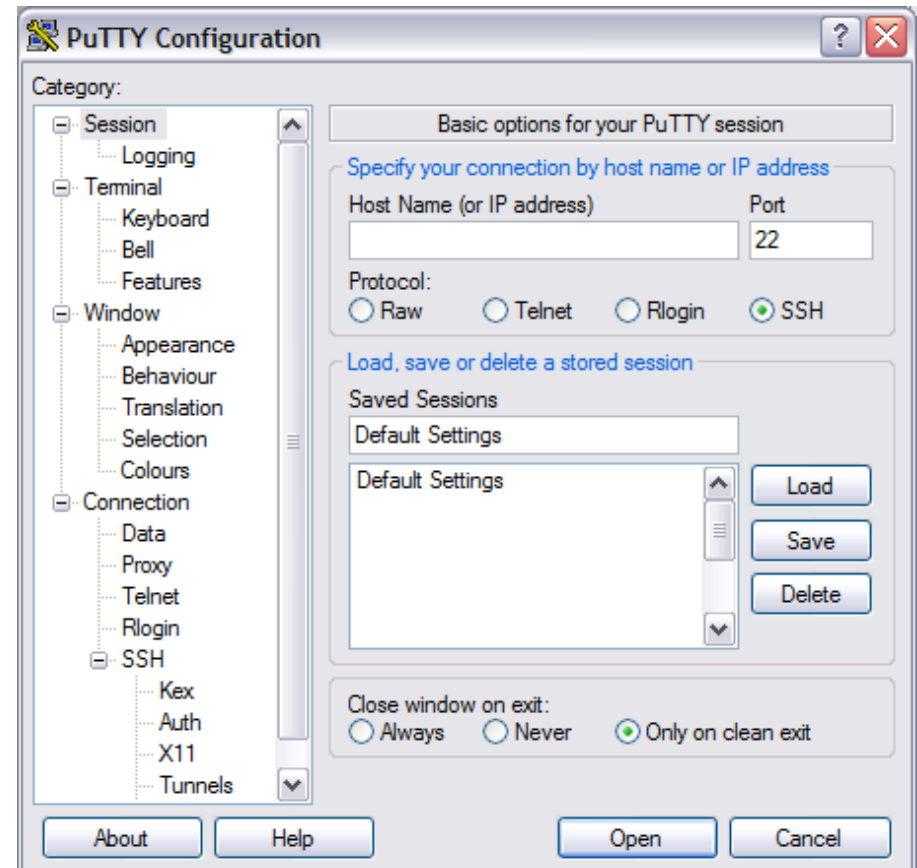
```
fedora login: _
```



Connect Over SSH Using Terminal.app or PuTTY.exe

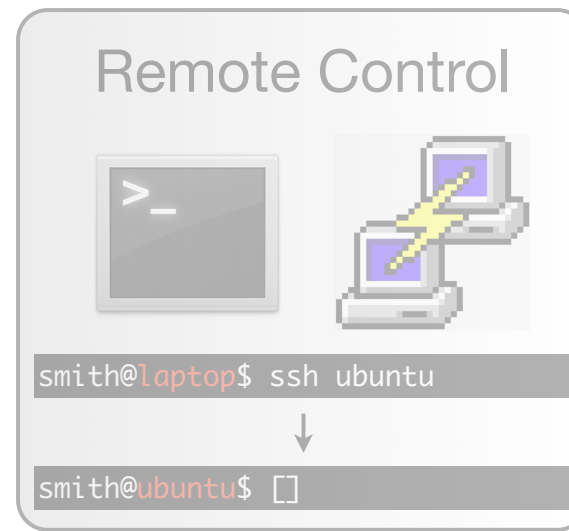


```
prion — bash — 53x17
Last login: Thu Aug 15 07:56:56 on ttys003
laptop:~ smith$
$ ssh workshop@192.168.56.101
workshop@192.168.56.101's password:
Welcome to Ubuntu 12.10 (GNU/Linux 3.5.0-36-generic i
* Documentation:  https://help.ubuntu.com/
New release '13.04' available.
Run 'do-release-upgrade' to upgrade to it.
Last login: Sun Aug 11 06:58:15 2013 from 192.168.56.
/home/workshop$
```



```
$ sudo shutdown -h now
```

Copy Files Over SSH Using scp or pscp



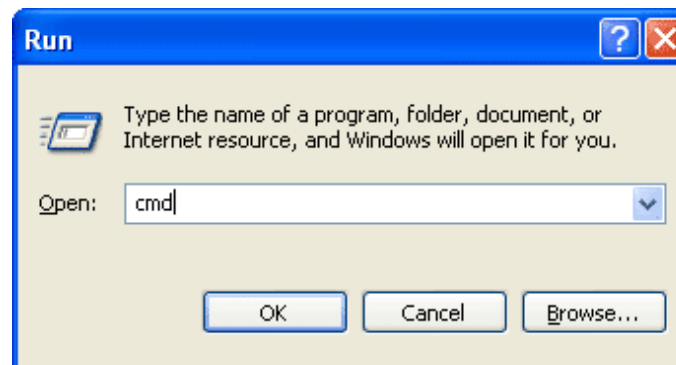
Mac

```
$ scp workshop@192.168.56.101:genomes/autism101.fa .
```

Windows

```
C:\Documents and Settings> cd "smith\My Documents\VirtualBox"
```

```
C:\ ... > pscp workshop@192.168.56.101:genomes/autism101.fa .
```



Working at a (Unix) Command Line Interface

cli, prompts, directories, paths, help, navigation, organizing files

Frederick J Tan
Bioinformatics Research Faculty
Carnegie Institution of Washington, Department of Embryology
tan@ciwemb.edu

26 August 2013

A Tale of Two Interfaces

Command Line

```
prion — ssh — 73x31
/home/workshop$ ls -l
total 24K
drwxr-xr-x 5 workshop workshop 4.0K Aug 19 22:20 autism
drwxrwxr-x 3 workshop workshop 4.0K Jul 25 13:47 bin
drwxr-xr-x 2 workshop workshop 4.0K Aug 18 17:44 genomes
drwxrwxr-x 3 workshop workshop 4.0K Jul 25 12:34 R
drwxr-xr-x 2 workshop workshop 4.0K Aug 19 19:14 splatter
drwxrwxr-x 6 workshop workshop 4.0K Jul 25 13:40 src
/home/workshop$ samtools sort

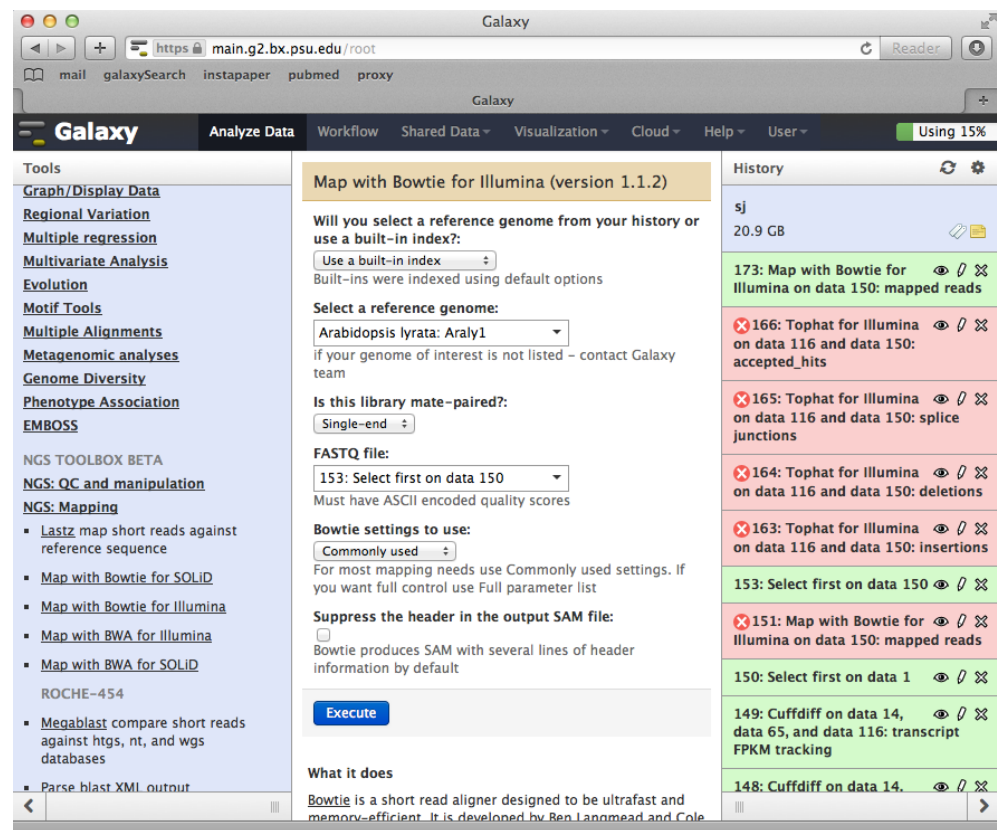
Usage:  samtools sort [options] <in.bam> <out.prefix>

Options: -n      sort by read name
         -f      use <out.prefix> as full file name instead of prefix
         -o      final output to stdout
         -l INT  compression level, from 0 to 9 [-1]
         -@ INT  number of sorting and compression threads [1]
         -m INT  max memory per thread; suffix K/M/G recognized [768M]

/home/workshop$
```

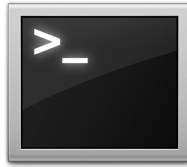
Terminal.app, Unix shell
breadth, cutting-edge

Graphical



Finder.app, Galaxy
discovery, visualization

Running Programs Using the Command-Line Interface



command-line



graphical

to run a program

type in the name of the
program and hit <ENTER>
(e.g. bowtie2)

double click on an icon

**to modify how a
program operates**

type in options after the program,
before hitting <ENTER>
(e.g. bowtie2 --very-sensitive)

click on check boxes,
select from pull-down menus

The Anatomy of a Shell Prompt

/home/workshop\$ □

The text before the dollar sign (\$) indicates what directory (folder) is currently “active”.

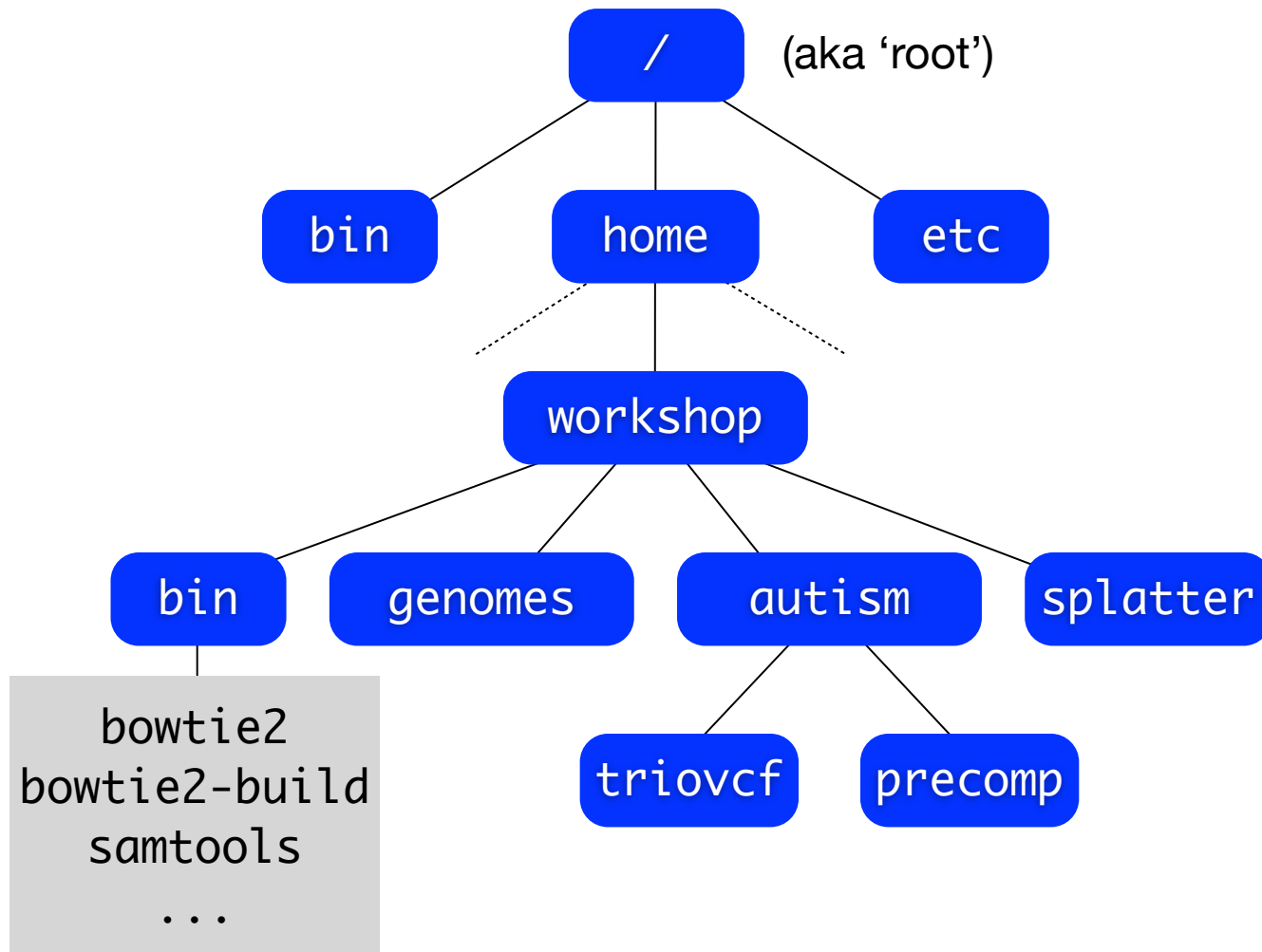
The \$ symbol indicates that the server is ready to perform a command.

The □ symbol indicates where what you type in will appear.

\$	Unix/Bash
C:\>	DOS/Windows
>>>	Python
>	R

DEFAULT : `workshop@server1:~$`

Unix Directory Structure



Task 1: List directory contents with `ls`

Shell Prompt `/home/workshop$` 

```
$ ls
```

```
$ ls bin
```

Color Coding

```
$ ls src
```

File Extensions

```
$ ls /home/workshop
```

```
$ ls /
```

```
$ ls bin/bowtie2-b*
```

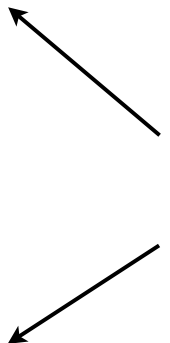
Wildcard

How to Get There from Here

Absolute

```
/home/workshop$ ls /home/workshop/bin
```

starts with a '/'
doesn't start with a '/'



Relative

```
/home/workshop$ ls bin
```

Task 2: Modify how `ls` works

```
$ ls -l
```

Manual Pages

```
$ man ls
```

(f)orward
(b)ack
(q)uit

- ♦ What does `ls` do?
- ♦ How does `-l` modify `ls`?
- ♦ How do you list **all** files?
i.e. don't ignore hidden files that begin with `'.'`

Divining Syntaxes - Example I

```
$ man ls
```

```
LS(1)
```

```
User Commands
```

```
LS(1)
```

```
NAME
```

```
ls - list directory contents
```

```
SYNOPSIS
```

```
ls [OPTION]... [FILE]...
```

↑
spaces are
very important

↑
0 or more
options
(optional!)

↑
0 or more
files

Try

```
$ ls
```

```
$ ls splatter
```

```
$ ls autism genomes
```

```
-a, --all
```

```
do not ignore entries starting with .
```

↑ ↑
single vs double dash

Divining Syntaxes - Example II

```
$ bowtie2
```

No index, query, or output file specified!

Bowtie 2 version 2.1.0 by Ben Langmead (langmea@cs.jhu.edu, www.cs.jh

Usage:

```
bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r>} [-S <sam>]
```

optional

'0 or more'

required

'OR'

Later we'll run

```
$ bowtie2 -x AUTISM101 -1 sample2_R1.fastq  
-2 sample2_R2.fastq -S sample2.sam
```

Task 3: Change current working directory

```
$ clear
```

```
$ pwd
```

```
$ ls autism/triovcf
```

```
$ cd autism
```

```
$ pwd
```

```
$ ls triovcf
```

```
$ cd triovcf
```

```
$ ls
```

```
$ ls .
```

```
$ ls ..
```

```
$ ls ~
```

```
$ pwd
```

```
$ ls
```

```
$ cd ../bin
```

```
$ cd ../../bin
```

```
$ ls
```

Path Resolution

Task 4: Figure out multiple ways to change to /home/workshop/src

Source, Destination; Source, Destination;
Source, Destination; Source, Destination;

```
$ man cp
```

NAME

cp - copy files and directories

SYNOPSIS

```
cp [OPTION]... [-T] SOURCE DEST
```

```
mv [OPTION]... [-T] SOURCE DEST
```

```
scp [-12346BCpqrsv] [-c cipher] [-F ssh_config] [-i identity_file]  
    [-l limit] [-o ssh_option] [-P port] [-S program]  
    [[user@]host1:]file1 ... [[user@]host2:]file2
```

WARNING

Unix assumes you know what you are doing.
Default settings will overwrite files if they exist.
There is no undo, Trash can, Recycle Bin, etc.

try using -i or -n flag

Task 5: Organize files into directories

```
$ cd
```

Defaults to \$HOME

```
$ mkdir newproject
```

```
$ mkdir newProject
```

Case Sensitive

```
$ touch newfile
```

```
$ ls
```

```
$ mv newfile newproject
```

```
$ ls
```

Task 6: Work more efficiently

```
$ touch asdf1hasdf1bjkasdfg
```

```
$ touch ziuymbciouyxcvbyv
```

```
$ mv a<TAB> n<TAB><TAB><TAB>P<TAB>
```

```
$ <UP><UP><LEFT><LEFT>...<LEFT>  
<BACKSPACE>{5}mv  
<CTRL>-E<SPACE>n<TAB>p<TAB>
```

Task 7: Remove files and directories

```
$ rmdir newProject
```

```
$ rm newProject/<TAB>
```

```
$ rmdir newProject
```

Document, Document, Document



130409-introToUnix-notes.rtf

Helvetica Bold 15

4/9/13

Start Ubuntu virtual machine instance

Start VirtualBox (v4.2.8)
... from <http://download.virtualbox.org/virtualbox/4.2.10/VirtualBox-4.2.10-84104-OSX.dmg>

Start Ubuntu 12.10 mini instance
... from <https://jshare.johnshopkins.edu/Ubuntu12.10-mini.zip>

Connect to Ubuntu instance

Open Terminal.app

ssh workshop@192.168.56.101

Task #1 and #2: top and man pages

top -i
... press 'q' to quit

man top
... press 'f' to forward, 'b' to back
... top shows list of running programs
... -i ignores programs that are idle (i.e. not running)!

Basic Unix Tasks and Commands

viewing files contents, redirecting pipes, editing with nano, searching

Frederick J Tan
Bioinformatics Research Faculty
Carnegie Institution of Washington, Department of Embryology
tan@ciwemb.edu

26 August 2013

Task 8: **View** 13known9unknown_rDNA.fasta **contents**

```
$ cd ~/splatter
```

```
$ head 13known9unknown_rDNA.fasta
```

```
$ cat 13known9unknown_rDNA.fasta
```

```
$ wc 13known9unknown_rDNA.fasta
```

```
$ tail 13known9unknown_rDNA.fasta
```

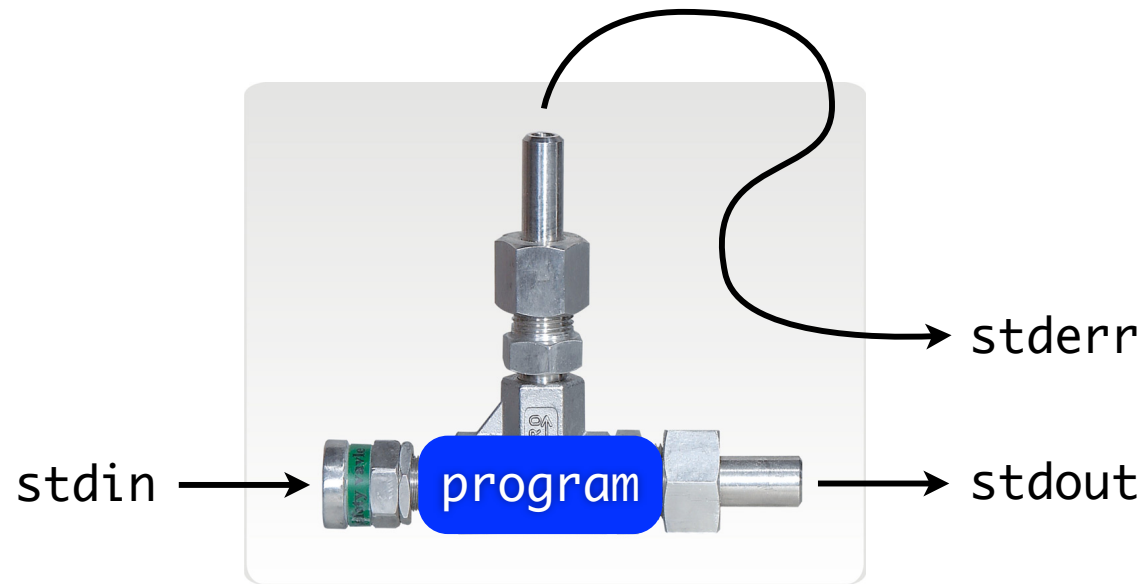
```
$ less 13known9unknown_rDNA.fasta
```

(f)orward
(b)ack
(q)uit

```
/Vibrio <ENTER>
```

(/)search

Streams, Pipelines, and Workflows



Redirect `$ head 13known9unknown_rDNA.fasta > first10.fa`

`$ less first10.fa`

Redirect (append) `$ head 13known9unknown_rDNA.fasta >> first10.fa`

`$ less first10.fa`

Pipe `$ head 13known9unknown_rDNA.fasta | less`

Task 9: Change file contents with nano text editor



TextWrangler, UltraEdit, Emacs, Vim, nano

<CTRL>-O
to save

<CTRL>-X
to quit

```
$ nano 13known9unknown_rDNA.fasta
```

```
GNU nano 2.2.6 File: 13known9unknown_rDNA.fasta
```

```
>Alcaligenaceae
AGTGAAATGCGTAGATATGCGGAGGAACACCGATGGCGAAGGCAGCCTCCTGGGATAACAC
TGACNNTCATGCACGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCC
CTGTAAACGATGTCAACTAGCTGTTGGGGCTTCGGGCTTGGTAGCGCAATAACGCGTGAAG
TTGACCGCCTGGGGAGTACGGTCGCAAGATTAACAACTCAAAGGAATTGACGGGGACCCGCA
CAAGCGGTGGATGATGTGGATTAATTCGATGCAACGCGAAAAACCTTACCTACCCCTTGACA
TGTCTGGAATCCTGAAGAGATTTAGGAGTGCTCGCAAGAGAACC GGAAACACAGGTGCTGCA
TGGCTGTCGTCAGCTCGTGTCTGTGAGATGTTGGGTAAAGTCCCGCAACGAGCGCAACCCCTT
GTCATTAGTT

>Pasteurellaceae
GGTGAAATGCGTAGAGATGTGGAGGAATACCGAAGGCGAAGGCAGCCCCCTTGGGAACGTAC
TGACGCTCATGTGCGAAAGCGTGGGGAGCAAACAGGATTAGATACCCTGGTAGTCCACGCT
GTAAACGCTGTCGATTTGGGGATTGGGCGATAAGCCTGGTGCCCCGAAGCTAACGTGATAAA
TCGACCGCCTGGGGAGTACGGCCGCAAGGTTAAACAACTCAAATGAATTGACGGGGGCCCGCA
CAAGCGGTGGAGCATGTGGTTTAATTCGATGCAACGCGAAGAACCTTACCTACTCTTGACA
TCCTCAGAATCCGGTAGAGATATCGGAGTGCCCTCGGGAACCTGAGAGACAGGTGCTGCATG
GCTGTCTGTCAGCTCGTGTGTGAAATGTTGGGTAAAGTCCCGCAACGAGCGCAACCCCTTAT
CCTTGTGTT

>Coxiellaceae
```

```
^G Get Help  ^O WriteOut  ^R Read File  ^Y Prev Page  ^K Cut Text   ^C Cur Pos
^X Exit       ^J Justify   ^W Where Is   ^V Next Page  ^U UnCut Text ^T To Spell
```

Task 10: Find all instances of AAA
in 13known9unknown_rDNA.fasta

```
$ grep AAA 13known9unknown_rDNA.fasta
```

Task 11: Figure out how to configure
grep to tell you....

How many lines contain AAA?

Which lines contain AAA?

What is the line preceeding the line with AAA?

How many lines do **not** have AAA?

Match Making with Regular Expressions

0 or more	<code>\$ grep "GT*AAAA" 13known9unknown_rDNA.fasta</code>
Wildcard	<code>\$ grep "G..AAAA" 13known9unknown_rDNA.fasta</code>
Repetition	<code>\$ grep "T\{1,5\}AAAA" 13known9unknown_rDNA.fasta</code>
'OR'	<code>\$ grep "[GCT]AAAA" 13known9unknown_rDNA.fasta</code>
Range	<code>\$ grep "[A-Z]AAAA" 13known9unknown_rDNA.fasta</code>
'NOT'	<code>\$ grep "[^G]AAAA" 13known9unknown_rDNA.fasta</code>
Start of line	<code>\$ grep "^AAA" 13known9unknown_rDNA.fasta</code>
End of line	<code>\$ grep "AAA\$" 13known9unknown_rDNA.fasta</code>

Task 12: Find all compressed files (.gz)

```
$ man find
```

```
$ find /home/workshop -name "*.gz"
```

Task 13: Figure out how to change working directory to place with sample2.tar.gz in 8 keys

Examining Deep Sequencing Data

FastQ format, quality control with FastQC

Frederick J Tan
Bioinformatics Research Faculty
Carnegie Institution of Washington, Department of Embryology
tan@ciwemb.edu

26 August 2013

FastQ Stores Both Sequence and Quality

```
$ cd ~/autism
$ ls
$ gunzip sample2.tar.gz
$ ls
$ tar xf sample2.tar
$ ls
$ less -S sample2_R1.fastq
```

- alternatively -
tar xzf sample2.tar.gz

press (q) to quit

FASTQ

```
@M01121:1:000000000-A2UB8:1:1101:15905:1864 1:N:0:1
GGTGCAGCAGTGGGCCCTGCTTGGCCGGGTCAGCTGGGCCGTCAGGCCT...
+
>1>11111>A11111AF00FF10AA000BA001D10////A/A/00AB...
```

← name
← sequence
← quality

worse !"#\$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ better
0.....26...31.....41

FASTA

```
>SequenceA
NGCGAAGTATATGGCACAATTGAGCAGTAAAAGTGTTAAAATATAG
```



Stringing Together Unix Commands Can Quickly Summarize Information

```
$ grep @M0 sample2_R1.fastq | cut -d: -f5 | uniq -c
```

7690	1101	7746	2101
7844	1102	8099	2102
7781	1103	7987	2103
7704	1104	8133	2104
7676	1105	8195	2105
7799	1106	8224	2106
7666	1107	7973	2107
7722	1108	8107	2108
7486	1109	7805	2109
7355	1110	7697	2110
7498	1111	7668	2111
7323	1112	7662	2112
7158	1113	7630	2113
7184	1114	7550	2114

14 tiles
top and bottom surfaces



Create a FastQC Quality Report

```
$ fastqc sample2_R1.fastq
```

```
Started analysis of sample2_R1.fastq  
Approx 5% complete for sample2_R1.fastq  
...  
Approx 100% complete for sample2_R1.fastq  
Analysis complete for sample2_R1.fastq
```

Open 192.168.56.101
in a web browser

FastQC Report

Summary

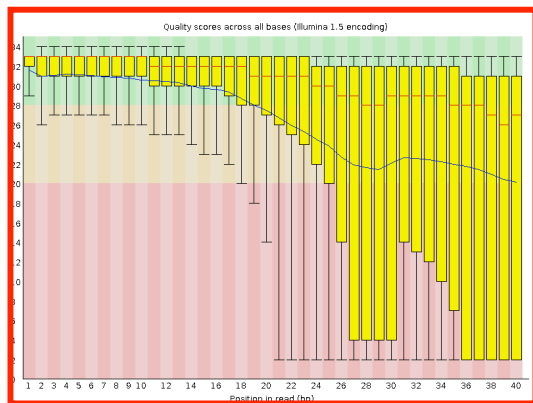
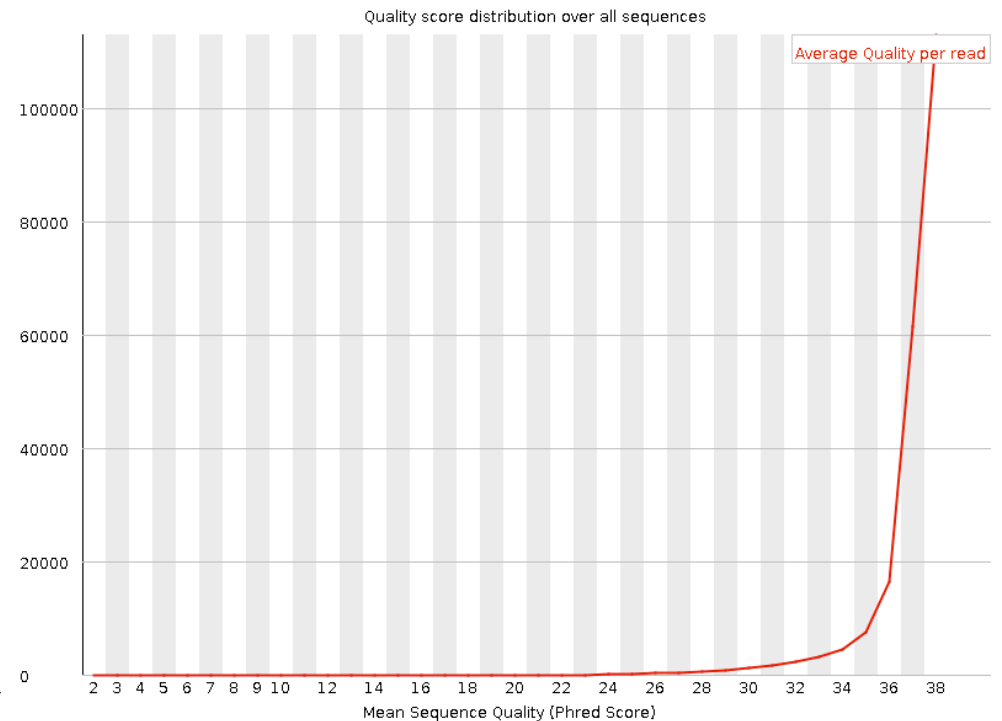
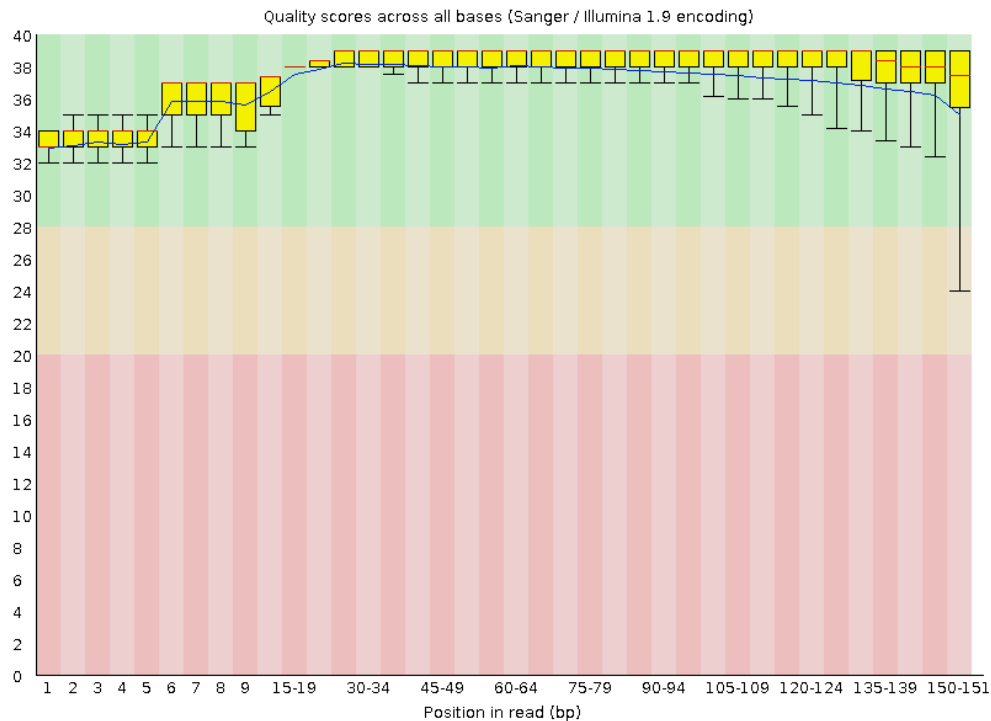
-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per base GC content](#)
-  [Per sequence GC content](#)
- 



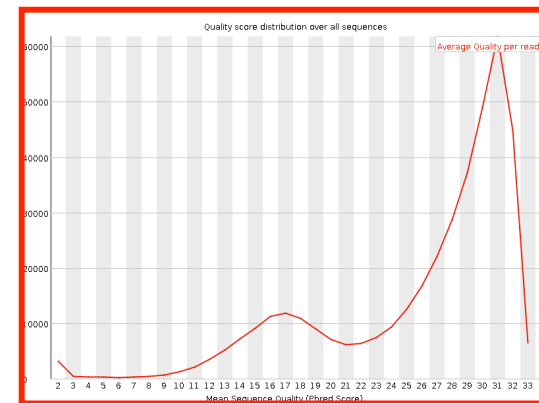
Basic Statistics

Measure	Value
Filename	sample2_R1.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	216362
Filtered Sequences	0
Sequence length	32-151
%GC	49

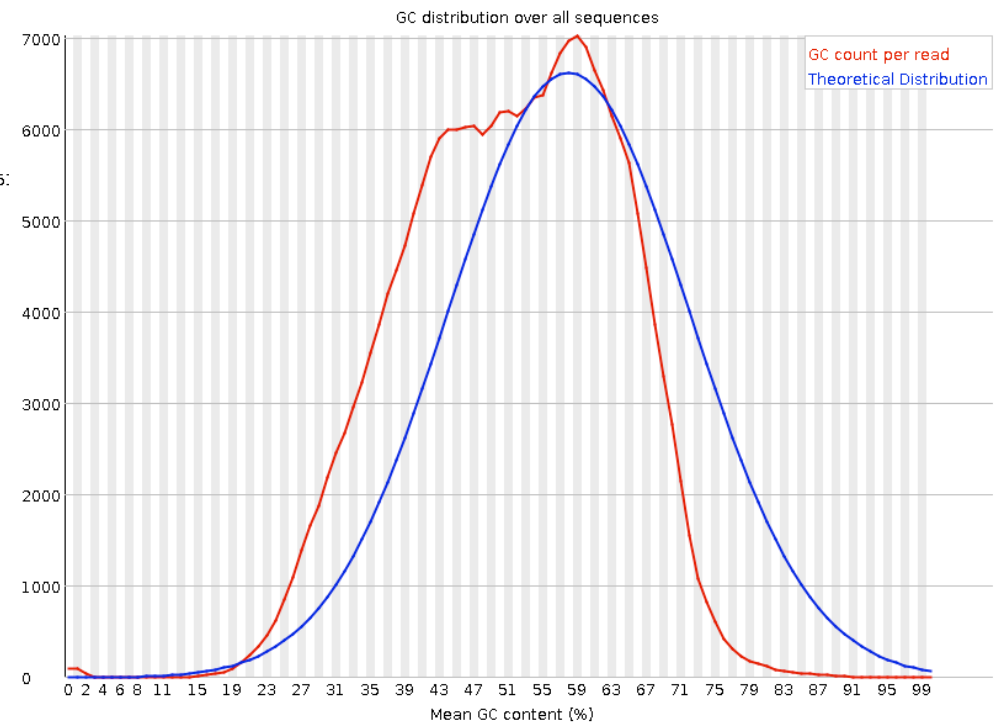
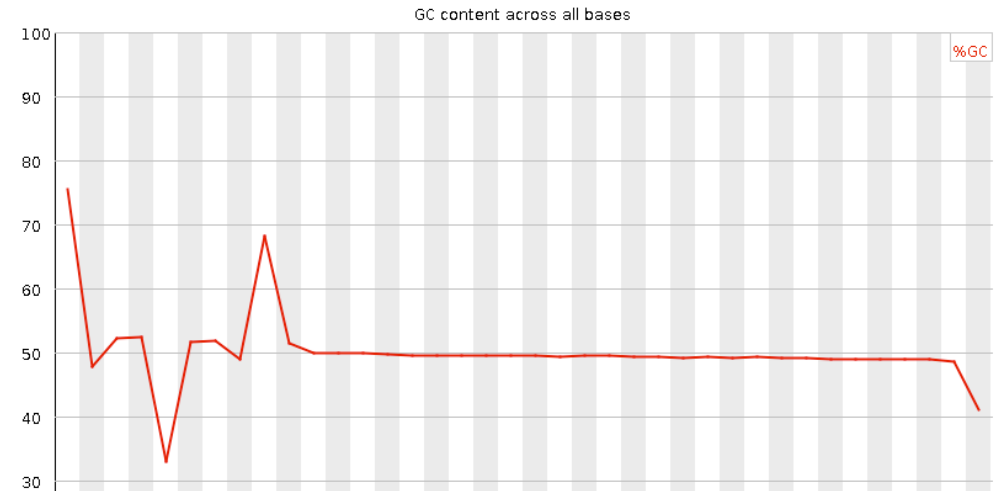
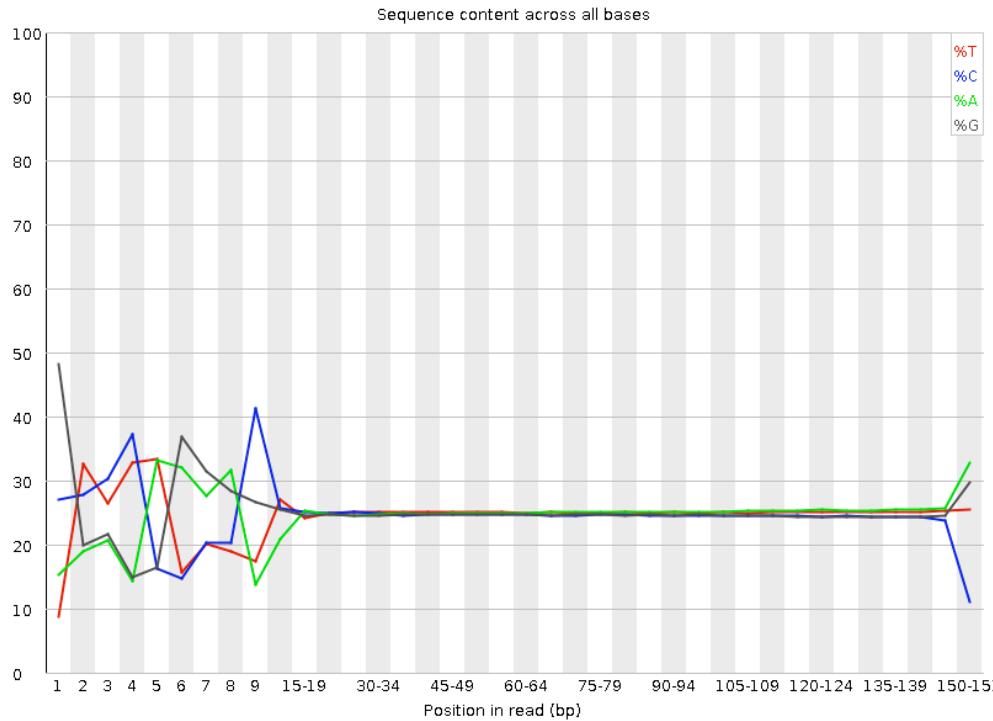
Quality Varies Both Across Position and Between Reads



worrisome

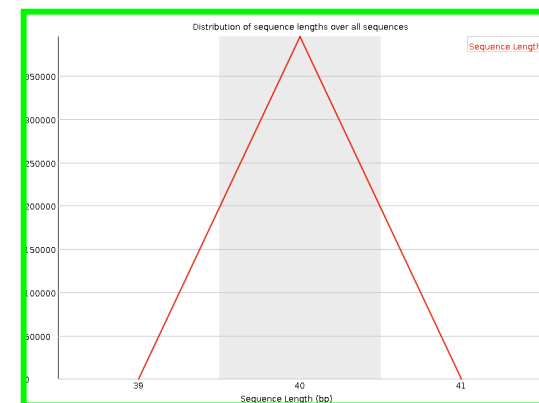
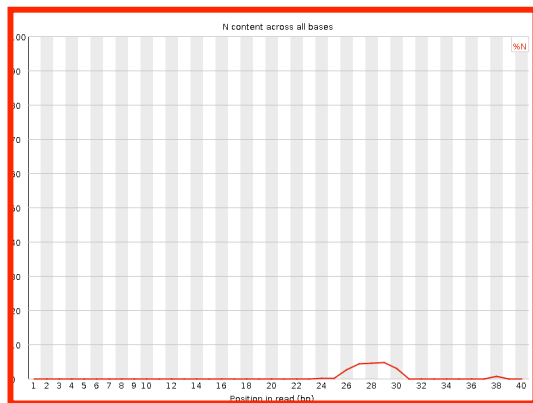
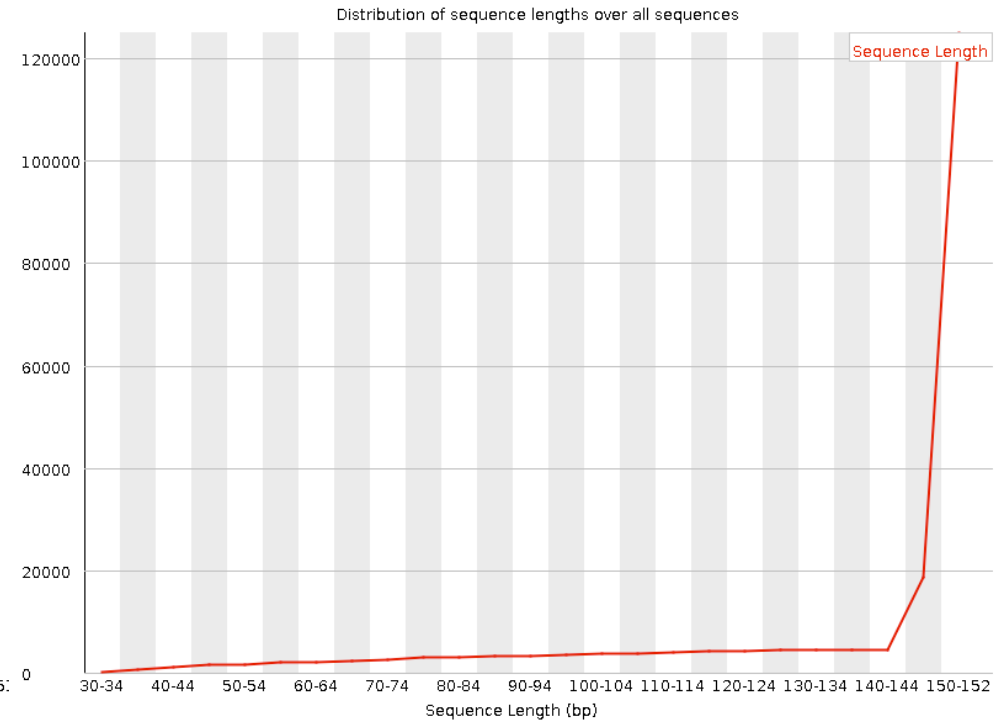
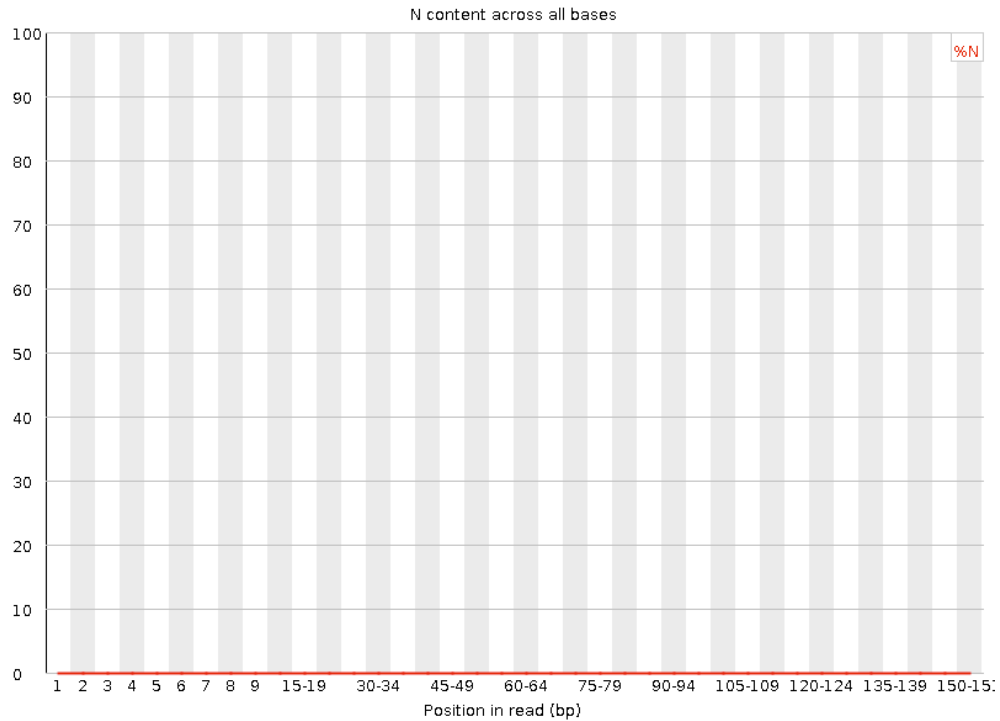


Sequence Composition Varies for Many Reasons



Adapter dimer
Library construction bias
Low complexity

N Content and Length Distributions Inform You About Your Run



Duplicate Sequences Are Summarized Multiple Ways

