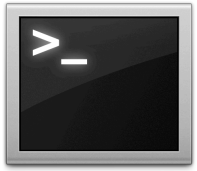
Quantitative Biology Bootcamp

Intro to Unix: Command Line Interface

Frederick J Tan
Bioinformatics Research Faculty
Carnegie Institution of Washington, Department of Embryology

2 September 2014

Running Programs Using the Command-Line Interface

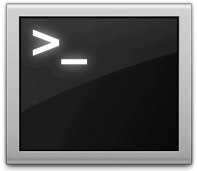
	 command-line	 graphical
strengths	breadth, cutting-edge	discovery, visualization
to run a program	type in the name of the program and hit <ENTER> (e.g. `bowtie2`)	double click on an icon
to modify how a program operates		

Task 1 : Run a simple program

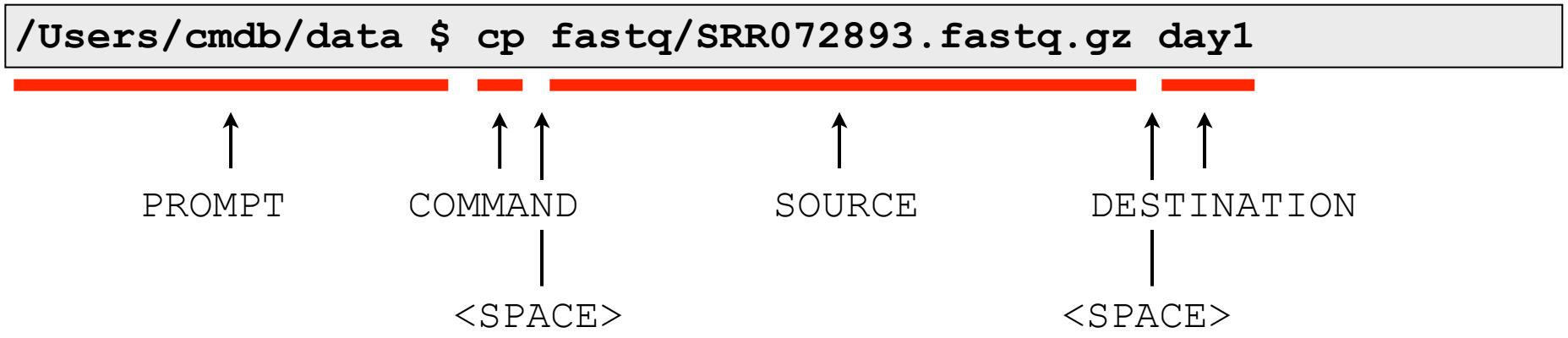
```
/Users/cmdb $ whoami
```

```
cmdb
```

Running Programs Using the Command-Line Interface

	 command-line	 graphical
strengths	breadth, cutting-edge	discovery, visualization
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

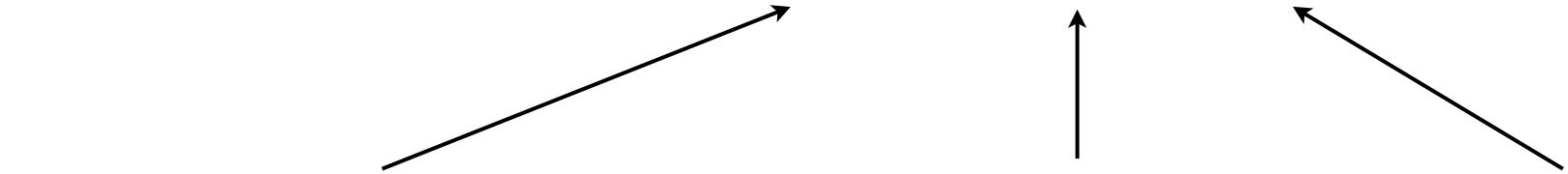
Breaking It Down With Spaces



The Anatomy of a Shell Prompt

```
/Users/cmdb/data $ cp fastq/SRR072893.fastq.gz day1
```

```
/Users/cmdb/data $
```

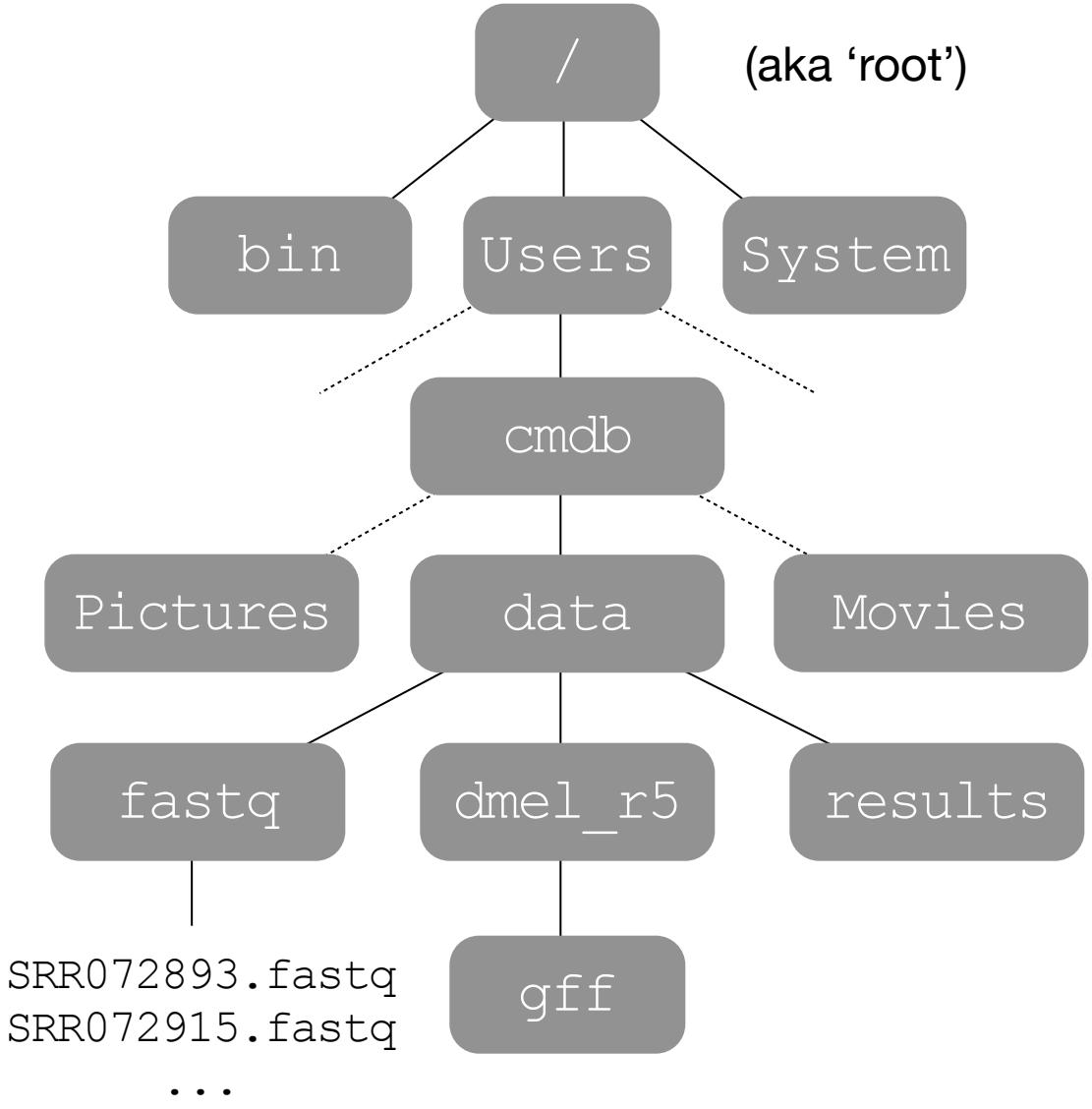


The text before the dollar sign ($) indicates the current working directory (i.e. "where you are")

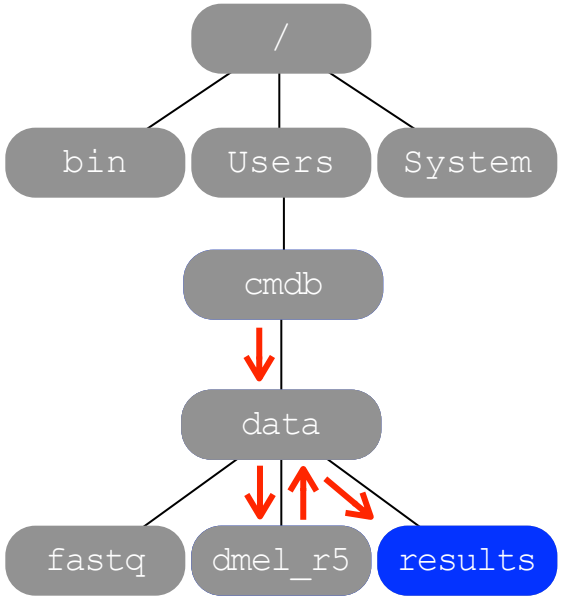
The $ symbol indicates that the server is ready to perform a command

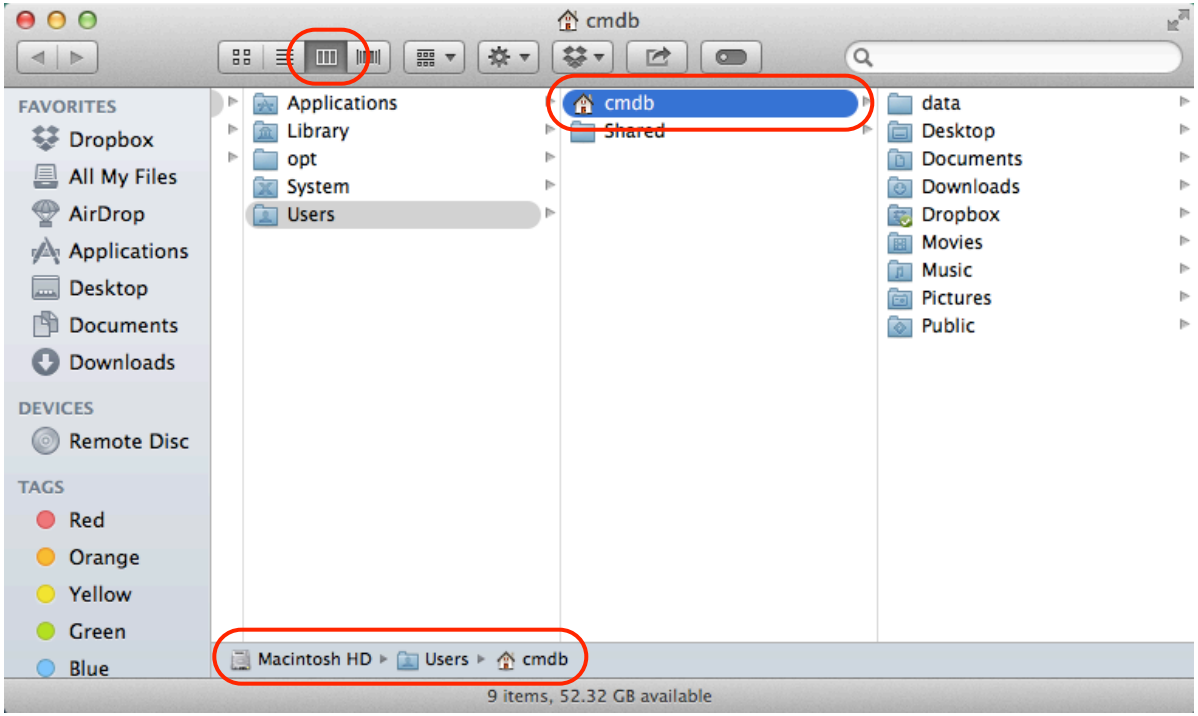
The ▯ symbol indicates where what you type in will appear

Unix Directory Structure

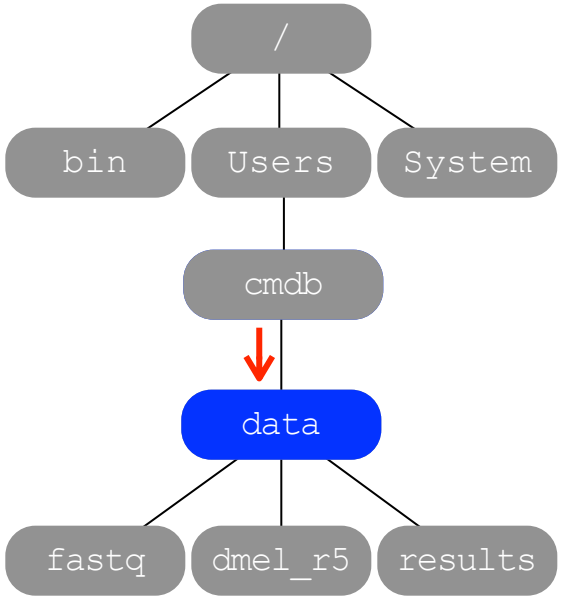


Task 2: Explore files using Finder.app





Task 3: Explore files using Terminal.app




```
/Users/cmdb $ ls
```

```
/Users/cmdb $ ls data
```

```
/Users/cmdb $ cd d<TAB>
```

<ENTER>

```
/Users/cmdb/data $ ls
```

<TAB>
<CTRL>
<UP>
<DOWN>

```
/Users/cmdb/data $ ls -l
```

```
/Users/cmdb/data $ mkdir day1
```

Options Options Everywhere

```
/Users/cmdb $ bowtie2 --very-sensitive dmel5 SRR072893.fastq
```

```
/Users/cmdb $ head -n 20 SRR072893.fastq
```

```
/Users/cmdb $ cut -d ':' -f 3 SRR072893.fastq
```

Quantitative Biology Bootcamp

Intro to Unix: First Commands and Finding Help

Frederick J Tan
Bioinformatics Research Faculty
Carnegie Institution of Washington, Department of Embryology

2 September 2014

Task 4: Copy SRR072893.fastq.gz into /Users/cmdb/data/day1

```
/Users/cmdb/data $ cp ???
```

```
/Users/cmdb/data $ man cp
```

```
CP(1)                     BSD General Commands Manual                    CP(1)

NAME
     cp -- copy files

SYNOPSIS
     cp [-R [-H | -L | -P]] [-fi | -n] [-apvX] source_file target_file
     cp [-R [-H | -L | -P]] [-fi | -n] [-apvX] source_file ...
        target_directory

DESCRIPTION
     In the first synopsis form, the cp utility copies the contents of the
     source_file to the target_file.  In the second synopsis form, the con-
     tents of each named source_file is copied to the destination
     target_directory.  The names of the files themselves are not changed.  If
     cp detects an attempt to copy a file to itself, the copy will fail.

     The following options are available:
```

(**f**)orward
(**b**)ack
(**q**)uit
(**h**)elp

```
/Users/cmdb/data $ cp fastq/SRR072893.fastq.gz day1
```

Task 5: Uncompress and view contents of .FASTQ file

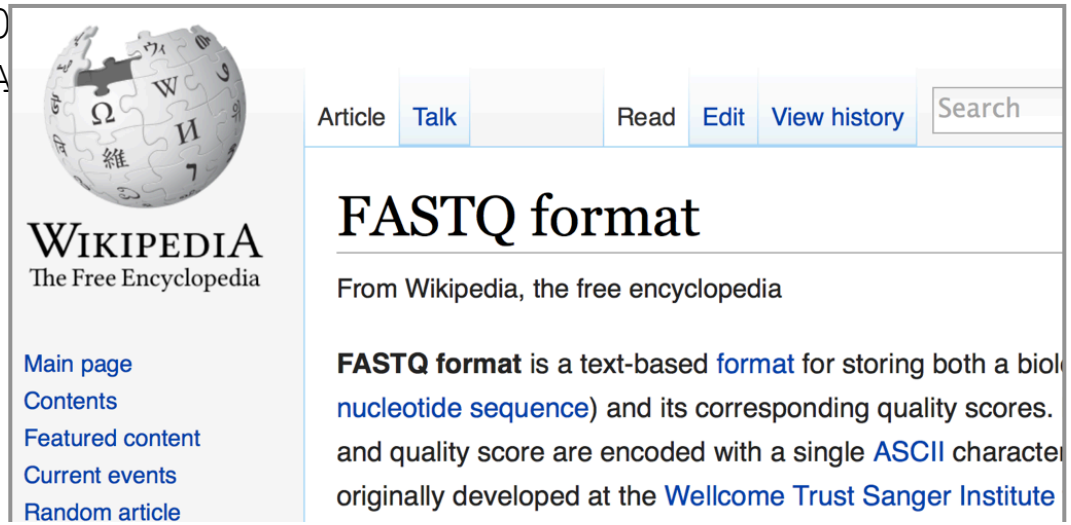
```
/Users/cmdb/data $ cd day1
```

```
/Users/cmdb/data/day1 $ gunzip S<TAB>
```

```
/Users/cmdb/data/day1 $ ls
```

```
/Users/cmdb/data/day1 $ head S<TAB>
```

```
@SRR072893.1 HWUSI-EAS585_0006:2:1:997:12409 length=40
NAATTATTCACCGATATCGCTTCAAGTGAACCCAAATAAT
+SRR072893.1 HWUSI-EAS585_0006:2:1:997:12409 length=40
!%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
@SRR072893.2 HWUSI-EAS585_0
NAGCAGCTGACCGAACTGAAGGGCAAA
```



FASTQ format Quality Scores

$$Q_{score} = -10 \log_{10} P$$

Probability of Sequencing Error	$\log_{10} P$	Quality Score	Encoding (Phred+33 / Sanger)
1%	-2	20	5
0.1%	-3	30	?
0.01%	-4	40	I

first 50 printable ASCII characters

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQR
```

20 30 40

```
@SRR925785.2
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAG
+
FDFFBFFFFAEFFDFE=EEEGGAGBFDFEBD=BC:B
```

Task 6: Interactively view .FASTQ file using `less`

```
/Users/cmdb/data/day1 $ less S<TAB>
```

```
@SRR072893.1 HWUSI-EAS585_0006:2:1:997:12409 length=40
NAATTATTCACCGATATCGCTTCAAGTGAACCCAAATAAT
+SRR072893.1 HWUSI-EAS585_0006:2:1:997:12409 length=40
!%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
@SRR072893.2 HWUSI-EAS585_0006:2:1:1000:19789 length=40
NAGCAGCTGACCGAACTGAAGGGCAAAAACTTCGAGAAGG
+SRR072893.2 HWUSI-EAS585_0006:2:1:1000:19789 length=40
!%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
@SRR072893.3 HWUSI-EAS585_0006:2:1:1006:15319 length=40
NGAACGCATGGAAATGAAGAAGCGCATCAAGAACTACCGC
+SRR072893.3 HWUSI-EAS585_0006:2:1:1006:15319 length=40
!%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
@SRR072893.4 HWUSI-EAS585_0006:2:1:1007:20582 length=40
NATGCACGCCTTTTCTATGCTCCCCATCTTGATTGATTCC
+SRR072893.4 HWUSI-EAS585_0006:2:1:1007:20582 length=40
!%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
@SRR072893.5 HWUSI-EAS585_0006:2:1:1008:11062 length=40
NTCAACAACAACAGACATTGATGATTTTCGGGGCTTTCGT
+SRR072893.5 HWUSI-EAS585_0006:2:1:1008:11062 length=40
!%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
SRR072893.fastq
```

```
(f)orward
(b)ack
(q)uit
(h)elp
```

Task 7: Open SRR072893.fastq in TextWrangler using `open`

```
/Users/cmdb/data/day1 $ open SRR072893.fastq ???
```

```
/Users/cmdb/data/day1 $ man open
```

```
OPEN(1)                   BSD General Commands Manual                  OPEN(1)

NAME
     open -- open files and directories

SYNOPSIS
     open [-e] [-t] [-f] [-F] [-W] [-R] [-n] [-g] [-h] [-b bundle_identifier]
          [-a application] file ... [--args arg1 ...]

DESCRIPTION
     The open command opens a file (or a directory or URL), [illegible] had
     double-clicked the file's icon. If no application name i[illegible]the
     default application as determined via LaunchServices is [illegible]the
```

```
(f)orward
(b)ack
(q)uit
(h)elp
```

```
/Users/cmdb/data/day1 $ open SRR072893.fastq -a TextWrangler
```

```
insufficient memory
```

Task 8: List file sizes using kilo/mega/giga bytes

```
/Users/cmdb/data/day1 $ ls -l
```

```
/Users/cmdb/data/day1 $ man ls
```

```
LS(1)                     BSD General Commands Manual                    LS(1)

NAME
     ls -- list directory contents

SYNOPSIS
     ls [-ABCFGHLOPRSTUW@abcdefghiklmnopqrstuwx1] [file ...]

DESCRIPTION
     For each operand that names a file of a type other than                s
     displays its name as well as any requested, associated                  For
     each operand that names a file of type directory, ls di                 mes
```

(**f**)orward
(**b**)ack
(**q**)uit
(**h**)elp
(**/**)search

```
/Users/cmdb/data/day1 $ ls -lh
```

Task 9: Count the number of reads

```
/Users/cmdb/data/day1 $ wc SRR072893.fastq
```

```
87571592 175143184 4582672018 SRR072893.fastq
```

```
/Users/cmdb/data/day1 $ man wc
```

```
WC(1)                          User Commands                          WC(1)

NAME
       wc - print newline, word, and byte counts for each file

SYNOPSIS
       wc [OPTION]... [FILE]...
       wc [OPTION]... --files0-from=F

DESCRIPTION
       Print newline, word, and byte counts for each FILE, and a total line if
       more than one FILE is specified.  With no FILE, or when FILE is -, read
       standard  input.   A  word  is a non-zero-length sequence of characters
       delimited by white space.  The options below  may  be  used  to  select
       which counts are printed, always in the following order: newline, word,
       character, byte, maximum line length.

       -c, --bytes
              print the byte counts

       -m, --chars
              print the character counts

 Manual page wc(1) line 1 (press h for help or q to quit)
```

```
(f)orward
(b)ack
(q)uit
(h)elp
```

```
/Users/cmdb/data/day1 $ wc -l SRR072893.fastq
```

```
87571592 SRR072893.fastq
```

How many reads?

Task 10: Display just the first million .FASTQ reads using `head`

```
$ head S<TAB>
```

```
$ man head
```

```
$ head -n4000000 S<TAB>
```

<CTRL>-C

<UP>

```
$ head -n4000000 SRR072893.fastq > SRR072893-1mil.fastq
```

Task 11: Open smaller file in TextWrangler

```
$ open S<TAB>-<TAB> -a TextWrangler
```

Quantitative Biology Bootcamp

Intro to Unix: grep, cut, uniq, and pipes

Frederick J Tan
Bioinformatics Research Faculty
Carnegie Institution of Washington, Department of Embryology

2 September 2014

Task 12: Find all occurrences of the pattern "GATTACA" using `grep`

```
$ grep "GATTACA" S<TAB>
```

<CTRL>-C

<UP>

```
$ grep "^GATTACA" SRR072893.fastq
```

Task 13: Color code each pattern match found using `grep`

```
$ man grep
```

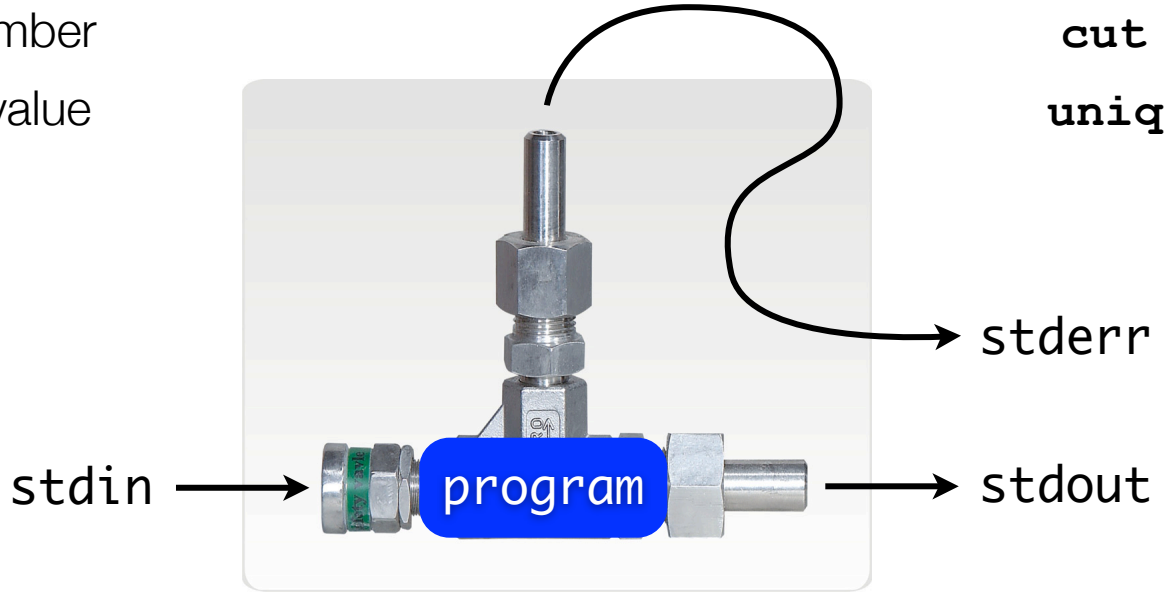
```
$ grep --color "GATTACA" SRR072893.fastq
```

Task 14: How many tiles are on an Illumina GAIIx flowcell?

```
$ head -n1 SRR072893.fastq
```

```
@SRR072903.1 HWUSI-EAS585_0006:2:1:997:12409 length=40
```

1. Extract sequence identifier — **grep**
2. Isolate tile number — **cut**
3. Find highest value — **uniq**




```
$ grep <PATTERN> SRR072893.fastq | cut <OPTIONS> | uniq
```

Task 14: How many tiles are on an Illumina GAIIx flowcell?

```
$ head -n1 SRR072893.fastq
```

```
@SRR072903.1 HWUSI-EAS585_0006:2:1:997:12409 length=40
```

1. Extract sequence identifier

```
$ grep "^@SRR" SRR072893.fastq
```

```
<CTRL>-C
<UP>
```

```
$ grep "^@SRR" SRR072893.fastq | head
```

2. Isolate tile number

```
$ grep "^@SRR" SRR072893.fastq | head | cut -d ':' -f3
```

3. Find highest value

```
$ grep "^@SRR" SRR072893.fastq | head | cut -d ':' -f3 | uniq
```

```
$ grep "^@SRR" SRR072893.fastq | cut -d ':' -f3 | uniq
```

Quantitative Biology Bootcamp

Intro to AWK

Frederick J Tan
Bioinformatics Research Faculty
Carnegie Institution of Washington, Department of Embryology

2 September 2014

AWK Splits Lines into Separate Fields

```
$ grep -v "^##" dmel-all-r5.57.gff | less -S
```

```
2L|FlyBase    |chromosome_arm |1|23011544|.  |.|.|ID=2L;Name=2L;Dbxref=REF
2L|FlyBase    |chromosome_band|1|1326937 |.  |+|.|ID=band-21_chromosome_ba
2L|FlyBase    |chromosome_band|1|22221   |.  |+|.|ID=band-21A5_chromosome_
2L|FlyBase    |chromosome_band|1|22221   |.  |+|.|ID=band-21A_chromosome_b
2L|sim4tandem|match          |1|457     |.  |+|.|ID=:3804126_sim4tandem_n
2L|sim4tandem|match_part     |1|457     |98 |+|.|Name=:9692572;Parent=:38
2L|FlyBase    |breakpoint     |1|1       |.  |.|.|ID=Df(2L)ED50001:bk1_bre
2L|FlyBase    |chromosome_band|1|1       |.  |+|.|ID=band-21A1_chromosome_
2L|FlyBase    |chromosome_band|1|1       |.  |+|.|ID=band-21A2_chromosome_
2L|FlyBase    |chromosome_band|1|1       |.  |+|.|ID=band-21A3_chromosome_
```

$1 **$2** **$3** **$4** **$5** **$6** **$7** **$8** **$9**

Quickly Filter and Manipulate Using AWK

```
$ awk '$3 == "TF_binding_site"' dmel-all-r5.57.gff
```

```
2L mE1_TFBS_HSA    TF_binding_site 4980 6482 . . . ID=FBsf00003
2L mE1_TFBS_cad    TF_binding_site 5126 6402 . . . ID=FBsf00002
2L BDTNP1_TFBS_dl  TF_binding_site 5203 6186 . . . ID=FBsf00002
2L BDTNP1_TFBS_hb  TF_binding_site 5203 6186 . . . ID=FBsf00002
2L BDTNP1_TFBS_Med TF_binding_site 5203 6186 . . . ID=FBsf00002
```

```
$ awk '{gsub(/;/, "\t"); print $9 "\t" $1 ":" $4 "-" $5}'
  dmel-all-r5.57.gff
```

```
ID=2L                                  2L:1-23011544
ID=band-21_chromosome_band             2L:1-1326937
ID=band-21A5_chromosome_band           2L:1-22221
ID=band-21A_chromosome_band            2L:1-22221
ID=:3804126_sim4tandem_na_gb.dmel      2L:1-457
```

```
$ awk '{total = total + $6} END {print total}' dmel-all-r5.57.gff
```

```
2.05551e+10
```