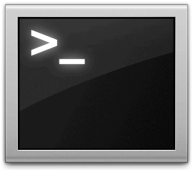
Open Your Favorite Note Taking Program

TextEdit.app

WordPad

Connect to Your Personal Bioinformatics Server



Start VirtualBox

Start PG2014

Minimize window

Connect via SSH using Terminal.app or PuTTY

View Reads Mapping to AZU1

Start IGV

Load exome.sorted.bam

Jump to AZU1 gene

Open a Web Browser

`etherpad.wikimedia.org`

PG2014

Practical GENOMICS

From Biology to Biostatistics

Monday, August 18th

8:15 AM	**Breakfast**
8:45 AM	**Workshop Overview**
9:30 AM	**Introduction to Unix and Exome Sequencing Part 1** Frederick Tan
12:00 PM	**Lunch**
1:00 PM	**Exome Sequencing Part 2** by Frederick Tan, Sarah Wheelan
3:00 PM	**Breakout Session**
5 PM	**Happy Hour**

Essential Day 1 Commands

```
$ cd genomes
```

```
$ bowtie2-build hg19.fa hg19
```

```
$ cd
```

```
$ mkdir day1
```

```
$ cd day1
```

```
$ bowtie2 -x ../genomes/hg19 -U ../rawdata/exome.fastq -S exome.sam
```

```
$ samtools view -bS exome.sam > exome.bam
```

```
$ samtools sort exome.bam exome.sorted
```

```
$ samtools index exome.sorted.bam
```

```
$ freebayes -f ../genomes/hg19.fa --pooled-discrete
  --pooled-continuous --min-alternate-fraction 0.05
  --genotype-qualities --report-genotype-likelihood-max
  exome.sorted.bam > exome.vcf
```

```
$ cp ~/answers/day1/snpEff.config .
```

```
$ java -jar ~/src/snpEff/snpEff.jar eff hg19 exome.vcf >
  exome.annotated.vcf
```

Practical Genomics: From Biology to Biostatistics

Introduction to Your Unix-Based Bioinformatics Server

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

18 August 2014

The Awesome Power of the Client-Server Model

Virtualization Provides Your Own Personal Bioinformatics Server

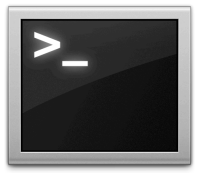
Physical Computer

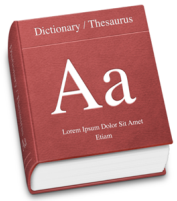
CPU

Memory

Hard Drive

VirtualBox





Virtual Computer

CPU

Memory

Hard Drive

Bioinformatics Server

ubuntu
linux for human beings

Analysis Programs

Datasets

SSH Allows Communication Between Client and Server

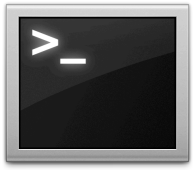
send commands

SSH connection

receive results


```
Welcome to Ubuntu 14.04.1 LTS (GNU/Linux 3.13.0-32-generic i686)

 * Documentation:  https://help.ubuntu.com/
Last login: Fri Aug  8 15:20:38 2014 from 192.168.56.1
/home/workshop $ []
```



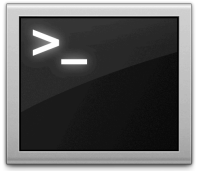
Terminal.app

Remote Control

putty.exe

www.chiark.greenend.org.uk/~sgtatham/putty

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

```
iMac:Desktop smith$ scp workshop@192.168.56.103:exome.bam* .
```

```
C:\Smith\Desktop > pscp workshop@192.168.56.103:exome.bam* .
```

The Anatomy of a Shell Prompt

```
iMac:Desktop smith$ scp workshop@192.168.56.103:exome.bam* .
```

```
C:\Smith\Desktop > pscp workshop@192.168.56.103:exome.bam* .
```

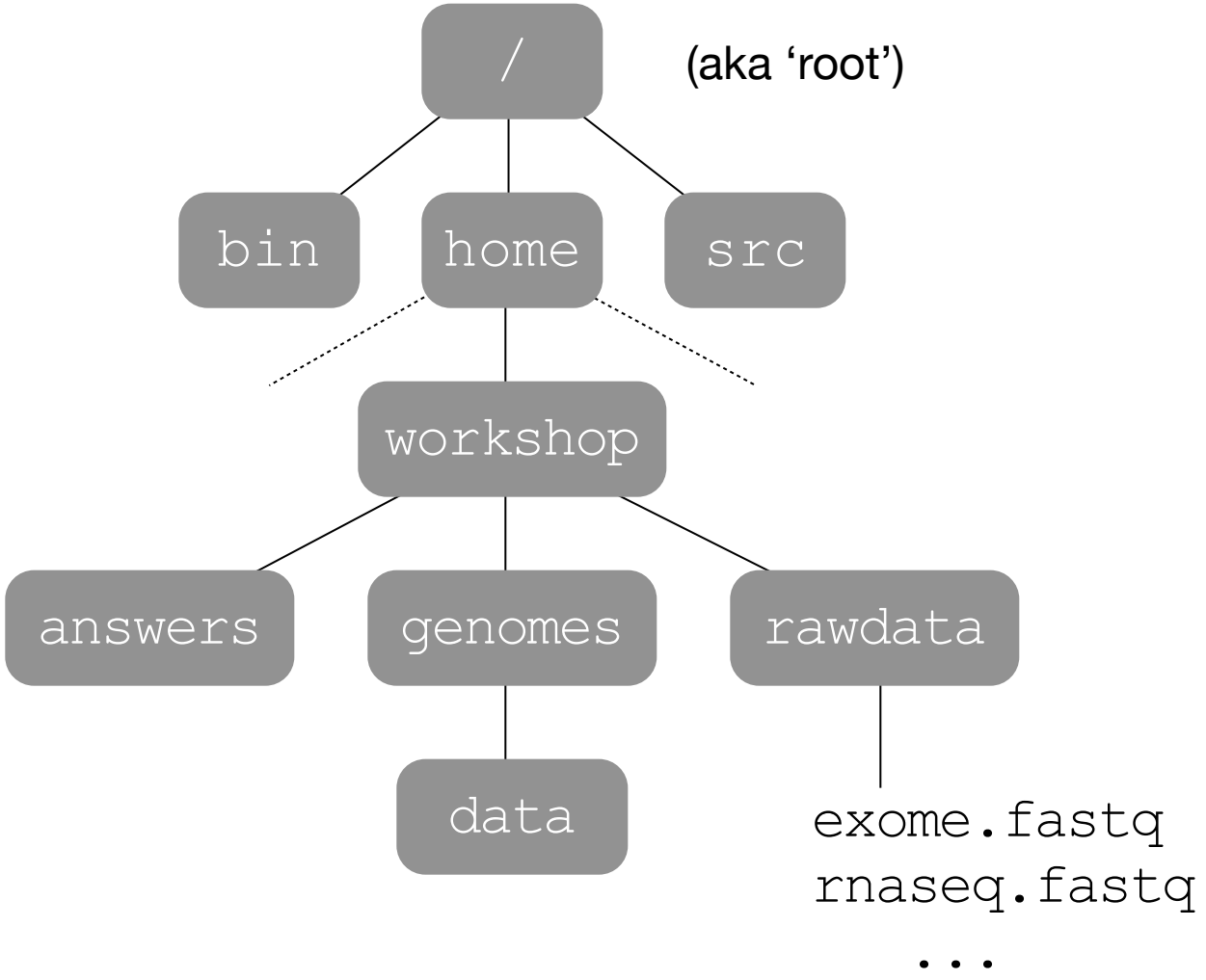
/home/workshop $ ▯

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

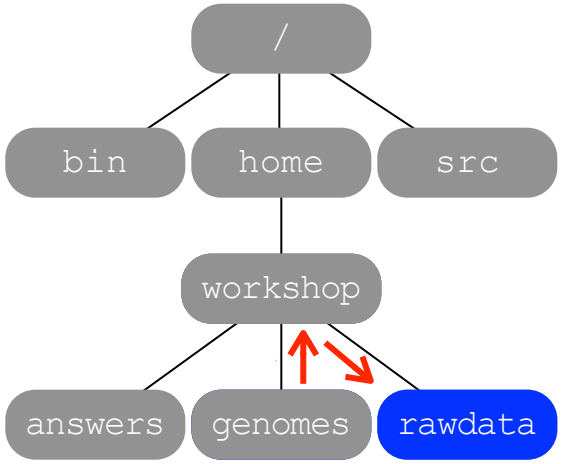
The ▯ symbol indicates where what you type in will appear

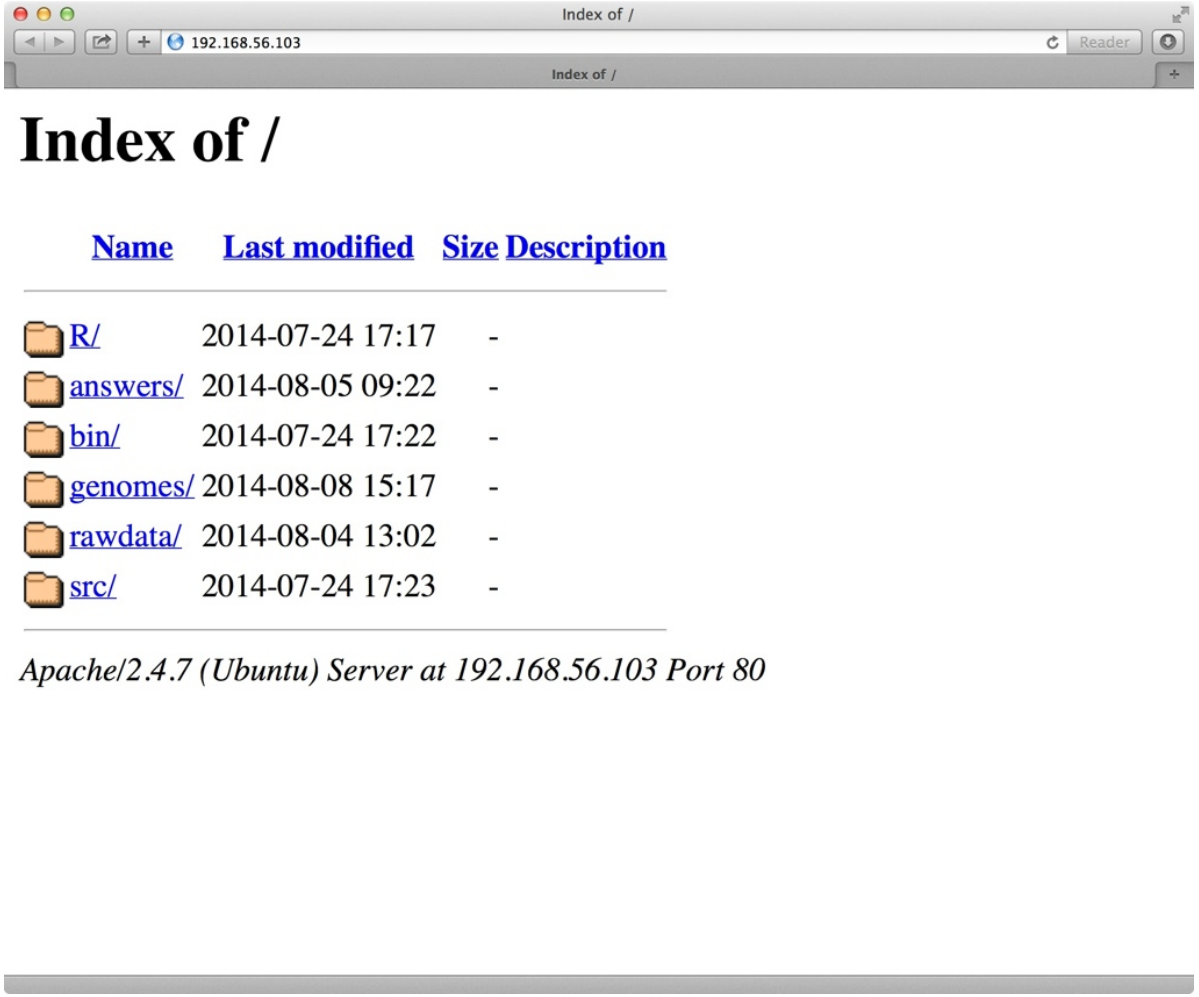
Unix Directory Structure



Task 1: Explore files on your server using a web browser

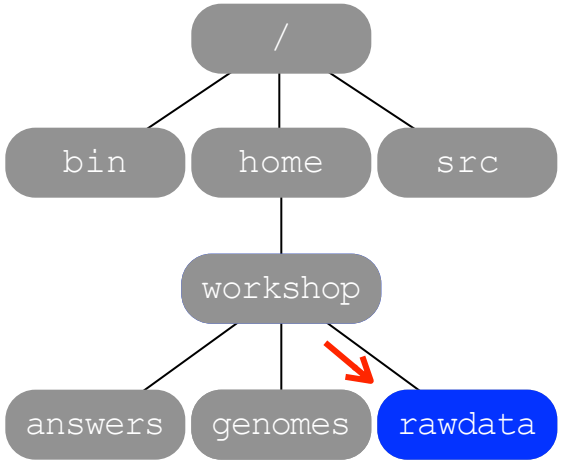
http://192.168.56.103





Task 2: Explore files on your server using the command line

ssh 192.168.56.103




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

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

```
/home/workshop $ cd r<TAB>
```

<ENTER>

<TAB>

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

<CTRL>

<UP>

```
/home/workshop/rawdata $ ls -l
```

<DOWN>

Practical Genomics: From Biology to Biostatistics

Exome-seq Analysis

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

18 August 2014

The Basics of Variant Calling

GOAL: Find mutations

Sequence subjects
(and "controls" like normal tissue or siblings)

Compare against reference genome

Prioritize putative variants

Followup with PCR, more subjects, genetic models, etc.

Whole Genome (Re)Sequencing

$3x10^9$ base genome

10^2 base read length

Q: How many reads does one need?

A: At least 30 million reads

Q: Problems?

1. Sequencing errors
2. Stochastic (Poisson)
3. Sequence Biases/Efficiency
4. Homozygous? heterozygous? heterogeneous?

Cost-Benefit Analysis

$$\text{Coverage} = \frac{\text{Number of } \textbf{reads} \text{ x Read } \textbf{length}}{\text{Size of } \textbf{genome}} \quad +/- \sqrt{\text{Coverage}}$$

$$f(k;\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

$$f(\mathbf{0};\lambda) = e^{-\lambda}$$

λ = events per interval

k = number of events

Coverage (λ)	1x	4.6x	15x	Exome (60x)
# of reads	30 million	138 million	450 million	54 million
Cost	$ 225	$1,000	$3,250	$400
$f(0;\lambda)$	1.1 billion	30 million	1 thousand	2.9 billion

The Art of Variant Calling

Does it segregate with phenotype?

What is the frequency in the population?

Expressed in "relevant" tissue(s)?

Effect on gene?

Important gene?

Exome-seq Analysis Pipeline

Quality Control Reads	FastQC	v0.11.2
Map Reads to Genome	Bowtie2	v2.1.0
Sort/Index Alignments	SAMtools	v0.1.19
Quality Control Alignments	IGV	v2.3.34
Identify Variants	FreeBayes	v0.9.14
Annotate Variants	SnpEff	v3.6c

hg19 (w/o alternate haplotypes)

UCSC knownGene Aug 2014

Raw Sequencing Data From Published Studies Are Available at SRA

NCBI Resources How To

SRA

SRA SRR925785

Save search Advanced

Display Settings: Full Send to:

SRX317792: GSM1172974: MCF10F Exome-Seq; Homo sapiens; OTHER
1 ILLUMINA (Illumina Genome Analyzer IIx) run: 52.5M spots, 10.6G bases, 4.6Gb downloads

chr19 subset

Accession: SRX317792
Experiment design: n/a
Submitted by: GEO
Study summary: SRP026538 • GSE48215: Exome sequencing of a panel of breast cancer cell lines to identify mutations •
All experiments • Run Selector (more...)
Sample: SAMN02225546 • MCF10F Exome-Seq (more...)
Library: (more...)
Platform: Illumina (more...)
Spot descriptor:

Experiment attributes:
GEO Accession: GSM1172974
Total: 1 run, 52.5M spots, 10.6G bases, 4.6Gb

#	Run	# of Spots	# of Bases	Si
1.	SRR925785	52,537,212	10.6G	

ID: 443327

Daemen *et al. Genome Biology* 2013, **14**:R110
http://genomebiology.com/2013/14/10/R110

Genome Biology

RESEARCH Open Access

Modeling precision treatment of breast cancer

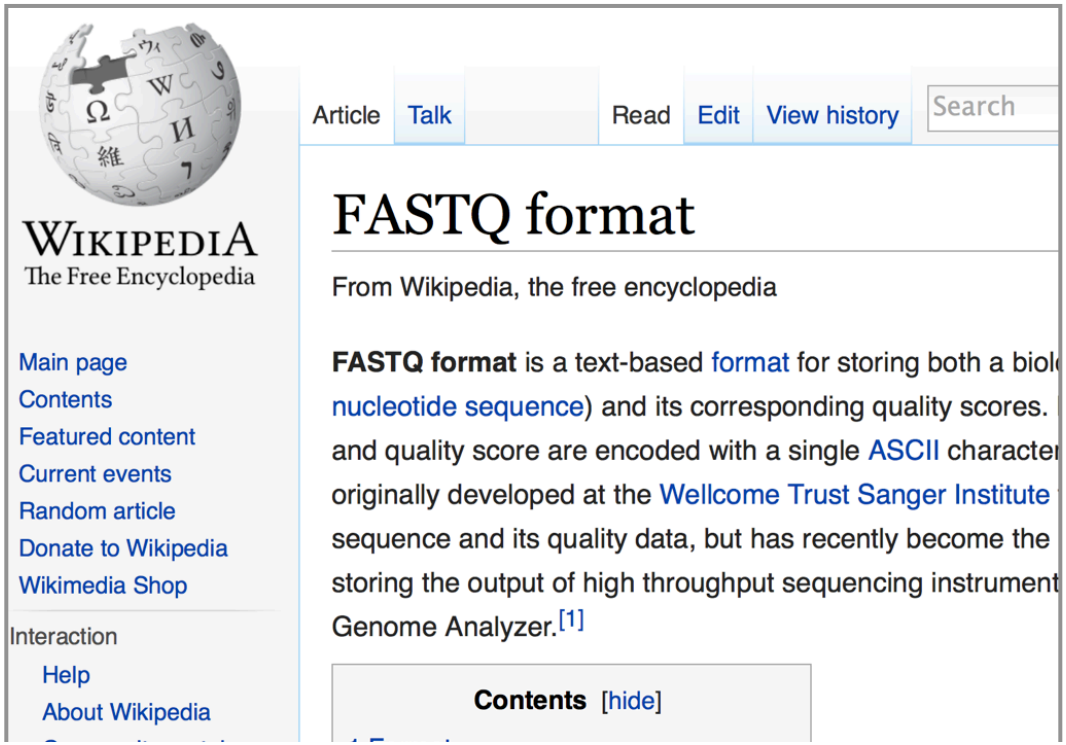
Anneleen Daemen[1,2,13*†], Obi L Griffith[1,3,6*†], Laura M Heiser[1,4], Nicholas J Wang[1,4], Oana M Enache[1], Zachary Sanborn[5], Francois Pepin[1,14], Steffen Durinck[1], James E Korkola[1,4], Malachi Griffith[6], Joe S Hur[7], Nam Huh[8], Jongsuk Chung[8], Leslie Cope[9], Mary Jo Fackler[9], Christopher Umbricht[9], Saraswati Sukumar[9], Pankaj Seth[10], Vikas P Sukhatme[10], Lakshmi R Jakkula[1], Yiling Lu[11], Gordon B Mills[11], Raymond J Cho[12], Eric A Collisson[1,2], Laura J van't Veer[2], Paul T Spellman[1,3] and Joe W Gray[1,4*]

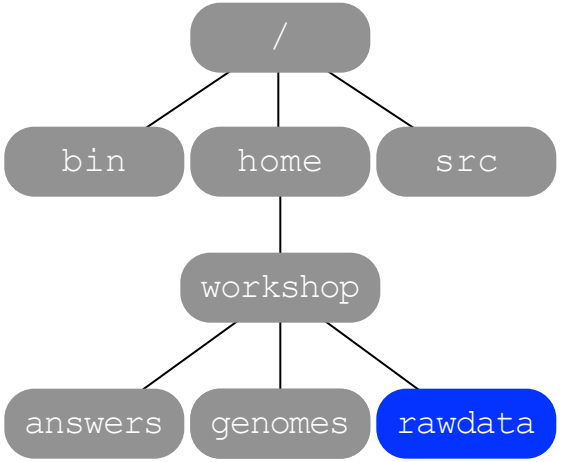
SRR925785 / pmid 24176112

Task 3: View contents of sequence reads

```
$ head exome.fastq
```

```
@SRR925785.2
GGTTAGGGTTAGGGTTAGGGTTAGGGTTAGGGTTAG
+
FDFFBFFFFAEFFDFE=EEEGGAGBFDFEBD=BC:B
@SRR925785.3
TGCGTTGTCCTCTGCACAGATTTCGGTGGTACTCTG
```





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 4: Count the number of reads

```
$ wc exome.fastq
```

```
 10881808  10881808 614641273 exome.fastq
```

```
$ 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
```

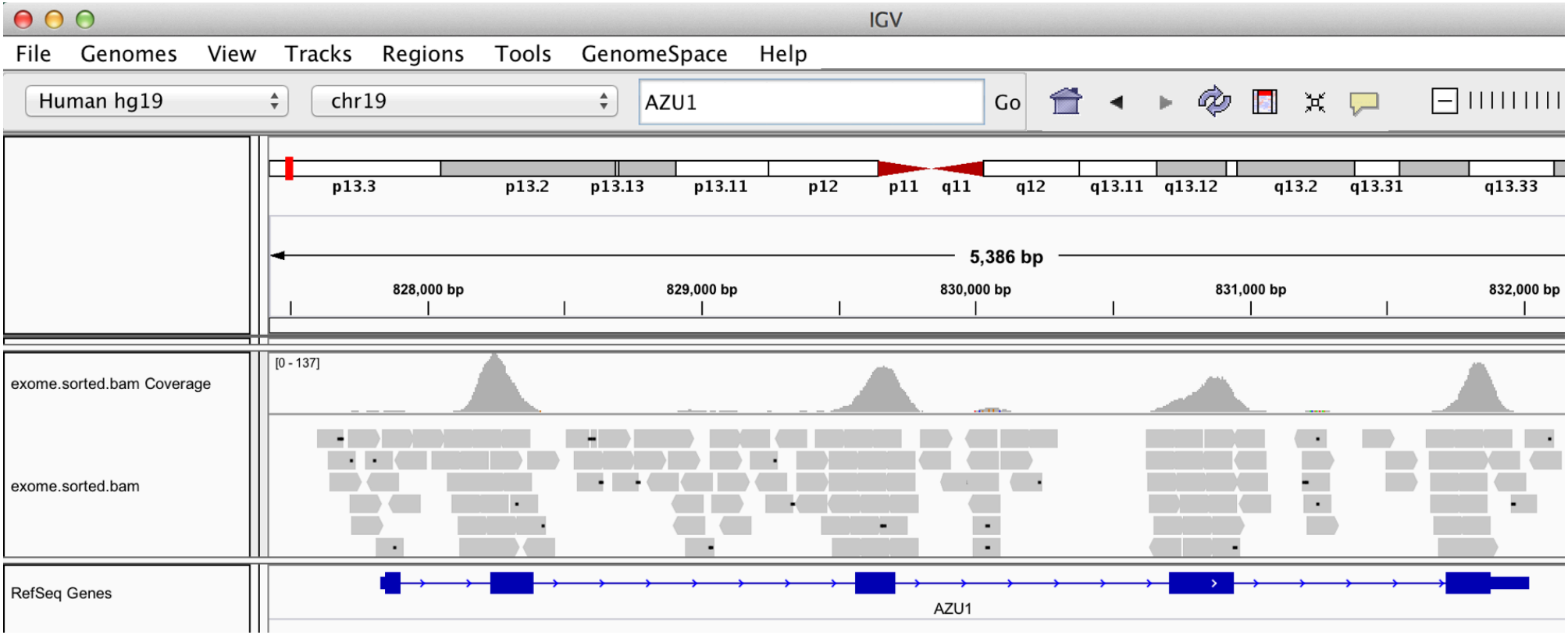
```
$ wc -l exome.fastq
```

```
10881808 exome.fastq
```

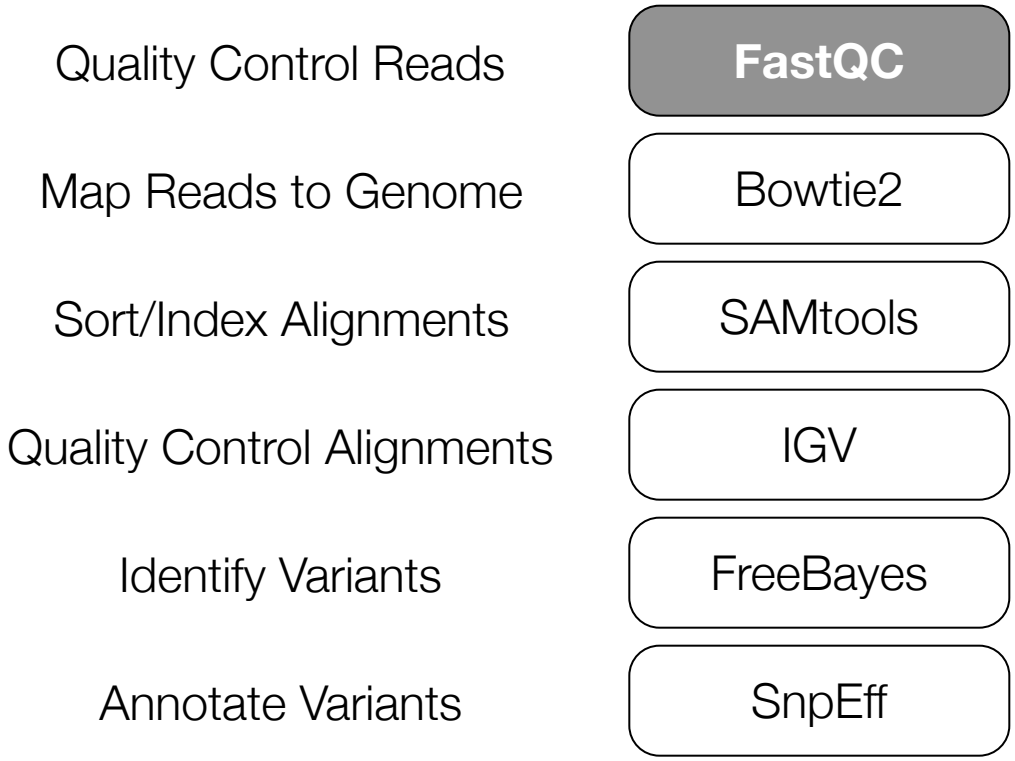
How many reads?

Task 5: Determine coverage of `exome.fastq` data set

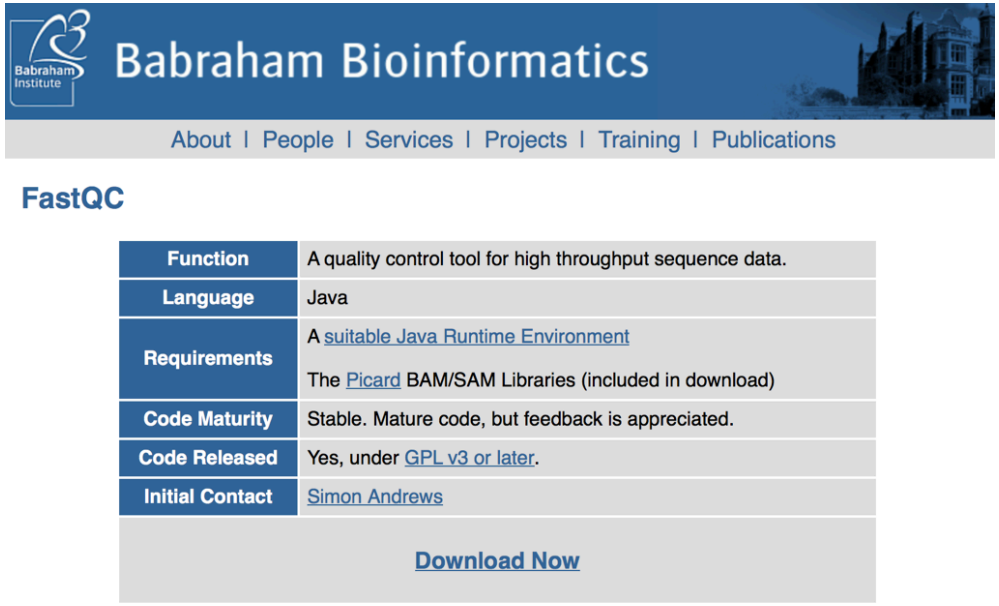
$$\frac{2.7 \times 10^6 \text{ reads} \times 101 \text{ bases}}{59 \times 10^6 \text{ bases in chr19}} = 4.6x \qquad \sim 7.8\% \text{ exonic} = 59x$$

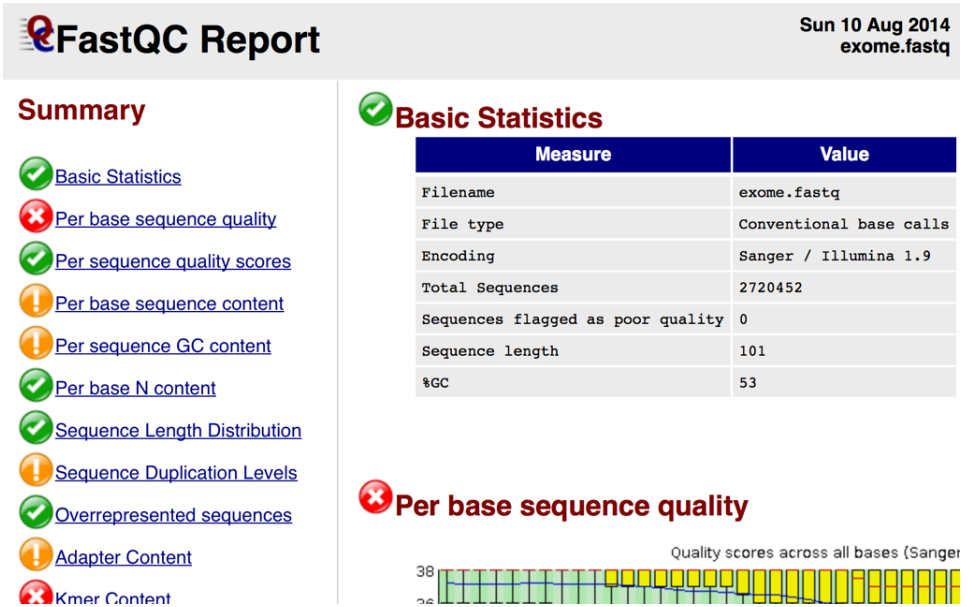


Exome-seq Analysis Pipeline



FastQC Provides Quality Control Reports on Sequence Reads





bioinformatics.babraham.ac.uk/projects/fastqc

Task 6: Summarize reads with FastQC

```
$ fastqc e<TAB>
```

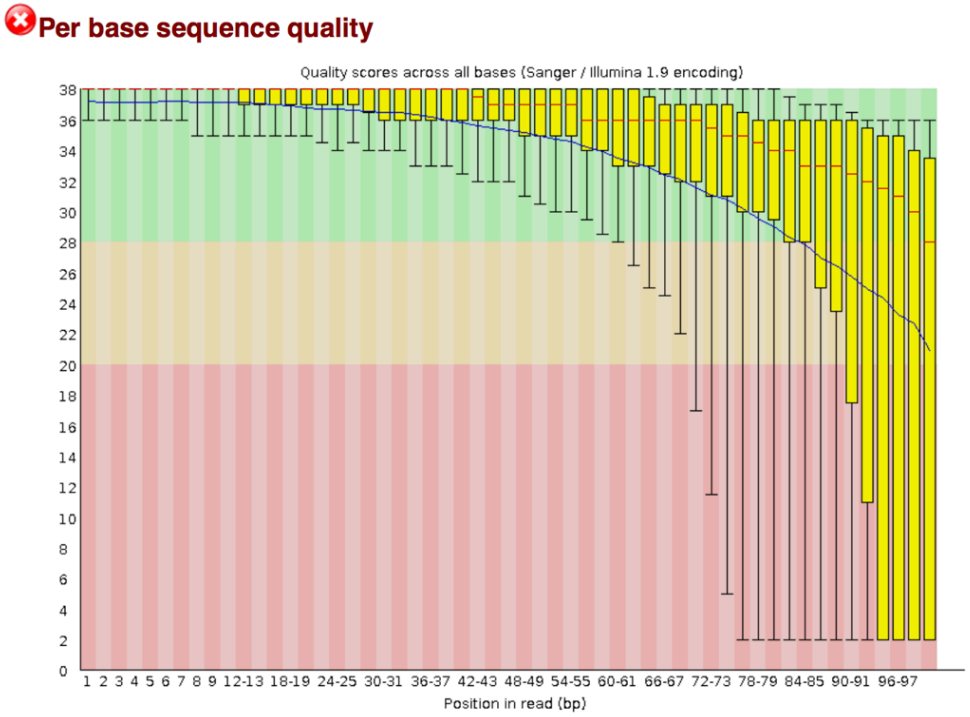
```
Started analysis of exome.fastq
Approx 5% complete for exome.fastq
Approx 10% complete for exome.fastq
Approx 15% complete for exome.fastq
                    ...
```

Basic Statistics

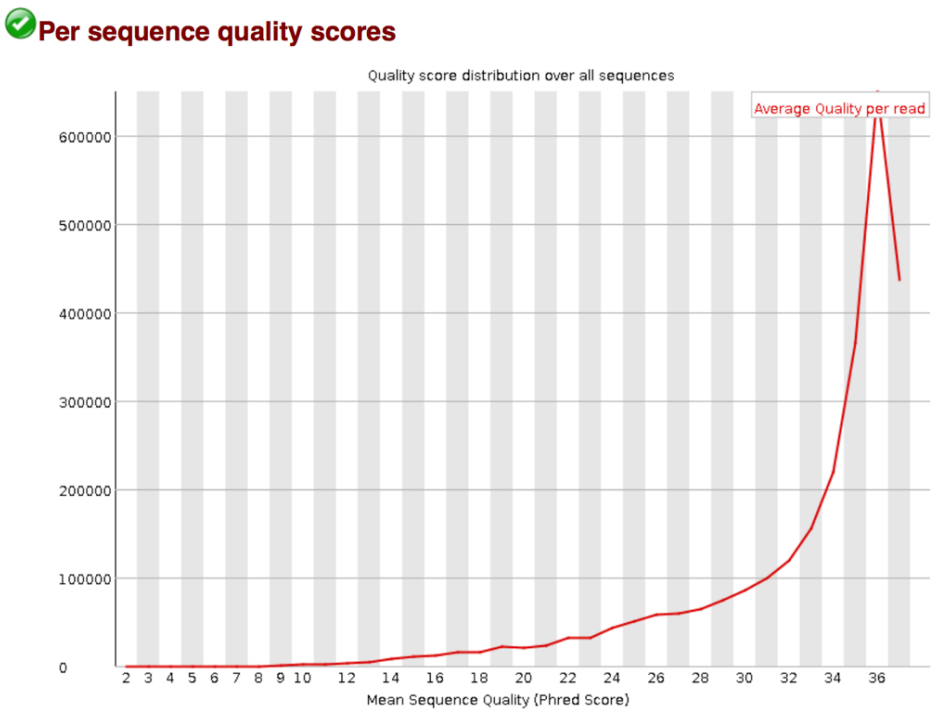
Measure	Value
Filename	exome.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2720452
Sequences flagged as poor quality	0
Sequence length	101
%GC	53

http://192.168.56.103/rawdata/exome_fastqc.html

Examine Quality on Both a Per Base and Per Sequence Basis



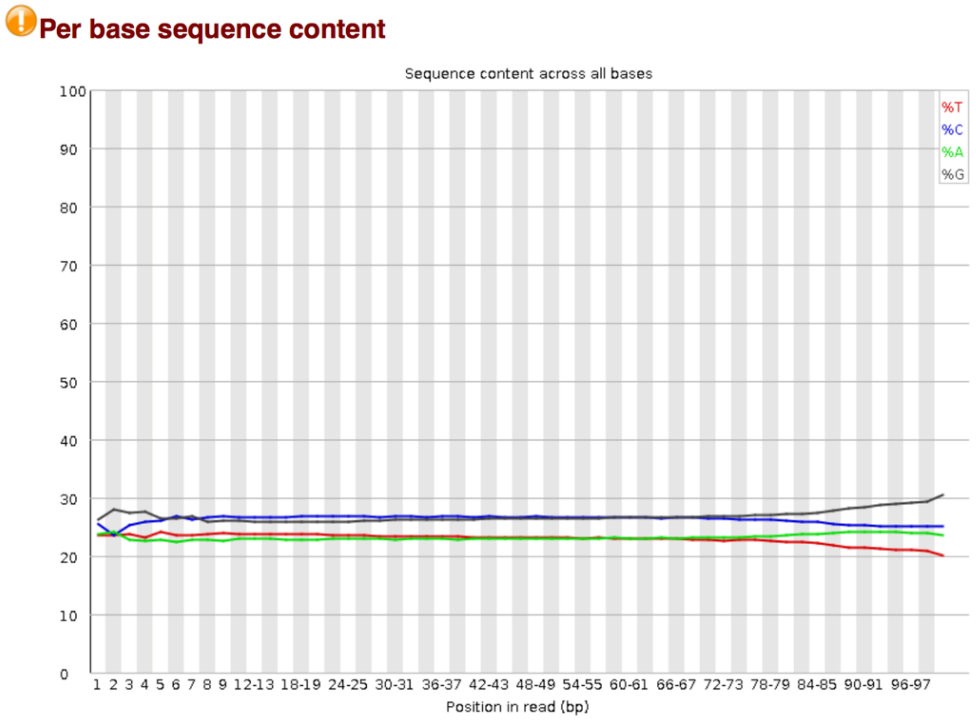
boxplot

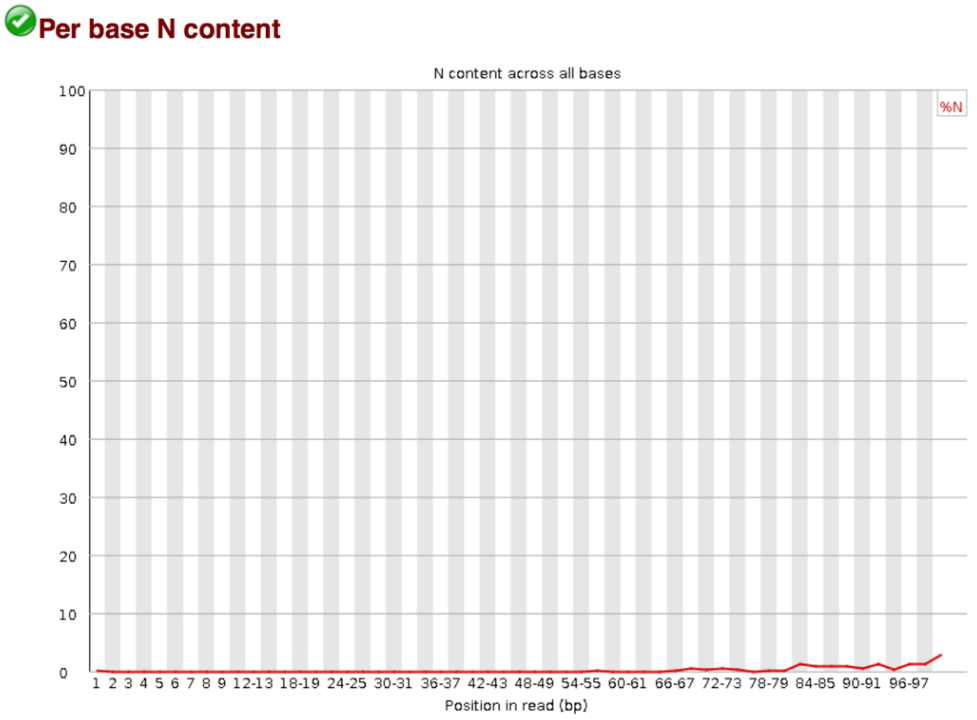


histogram

read too long
bad portion of flowcell

Quality Control Statistics Generally Provide Hints and Require Interpretation

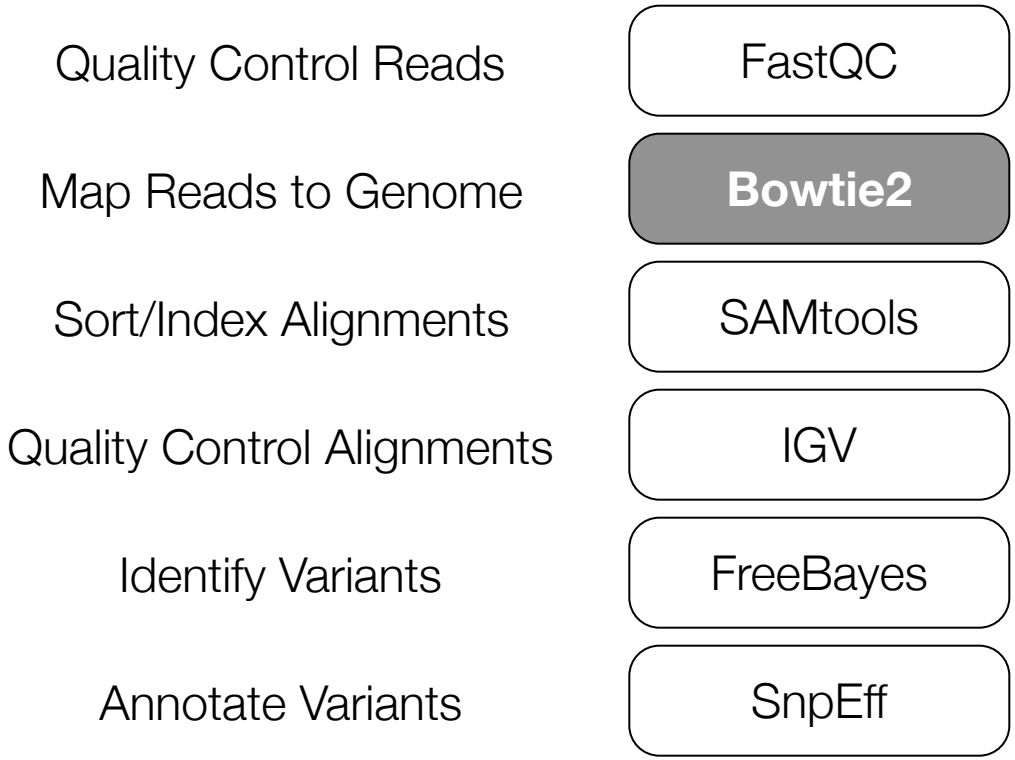




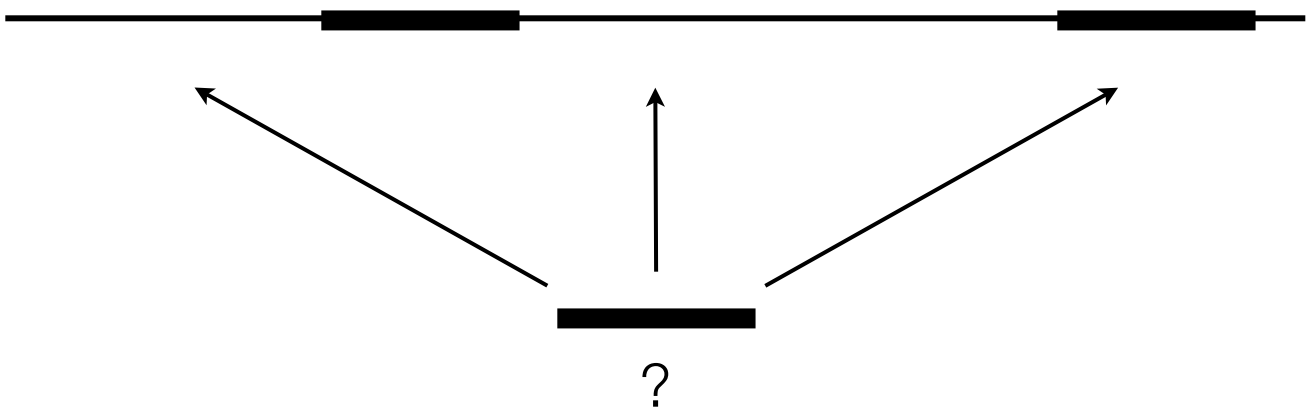
adapter
over-represented sequence (rRNA, adapter dimer)
bias (fragmentation, priming, ligation)

bad cycle
(focus, diversity)

Exome-seq Analysis Pipeline



The Mapping Problem



GOALS

Find match quickly

Find true match

Avoid missing a match

COMPLICATIONS

Large genome, many reads

Paralogs, pseudo-genes, repeats

Sequencing errors, polymorphisms

Mapping Solutions

<CTRL>-F **BLAST** **BLAT** **Bowtie2**

how many sequences do you need to search?

hundreds or billions

how many genomes are you searching against?

one or many

how many mismatches will you consider?

few or many

An "ultrafast and memory-efficient tool" for aligning reads to a genome

Table of Contents

Version **2.2.2**

billions of reads
(50 - 1000+ bases)

gapped and local alignments

reports a single alignment (by default)

bowtie-bio.sourceforge.net/bowtie2/manual.shtml

Cheatsheet for `bowtie2` Syntax

```
$ man bowtie2
```

```
No manual entry for bowtie2
See 'man 7 undocumented' for help when manual pages are not available.
```

```
$ bowtie2
```

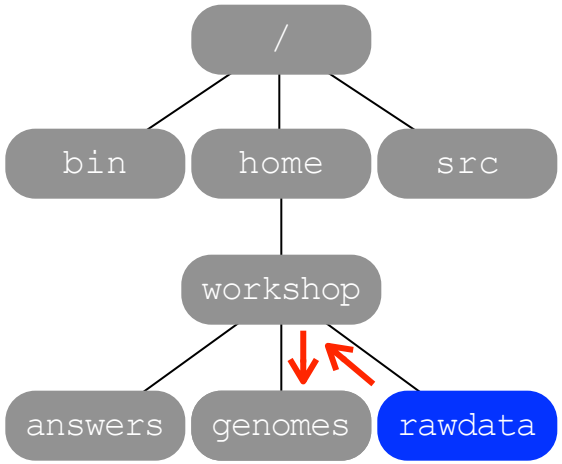
```
No index, query, or output file specified!
Bowtie 2 version 2.1.0 by Ben Langmead (langmea@cs.jhu.edu, www.cs.jhu.edu/~langmea)
Usage:
  bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r>} [-S <sam>]

  <bt2-idx>  Index filename prefix (minus trailing .X.bt2).
             NOTE: Bowtie 1 and Bowtie 2 indexes are not compatible.
  <m1>       Files with #1 mates, paired with files in <m2>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <m2>       Files with #2 mates, paired with files in <m1>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <r>        Files with unpaired reads.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <sam>      File for SAM output (default: stdout)
```

Task 7: Change working directory to `/home/workshop/genomes`

```
/home/workshop/rawdata $ cd ..
```

```
/home/workshop $ cd genomes
```




```
/home/workshop/rawdata $ cd ../genomes
```

```
/home/workshop/rawdata $ cd /home/workshop/genomes
```

```
/home/workshop/rawdata $ cd ~/genomes
```

↑

~ is a nickname for `/home/workshop`

Task 8: Examine contents of `hg19.fa`

```
$ ls
```

```
$ head hg19.fa
```

```
$ less hg19.fa
```

```
>chr19
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
```

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

```
$ grep "chr" hg19.fa
```

```
$ grep "GATTACA" hg19.fa
```

```
$ grep -n "GATTACA" hg19.fa
```

How many lines have GATTACA?

Task 9: Build a Bowtie2 genome index

```
$ bowtie2-build
```

```
No input sequence or sequence file specified!
Bowtie 2 version 2.1.0 by Ben Langmead (langmea@cs.jhu.edu, www.cs.jhu.edu/~langmea)
Usage: bowtie2-build [options]* <reference_in> <bt2_index_base>
    reference_in            comma-separated list of files with ref sequences
    bt2_index_base          write .bt2 data to files with this dir/basename
```

```
$ bowtie2-build hg19.fa hg19
```

Task 10: Map reads with Bowtie2

```
$ cd
```

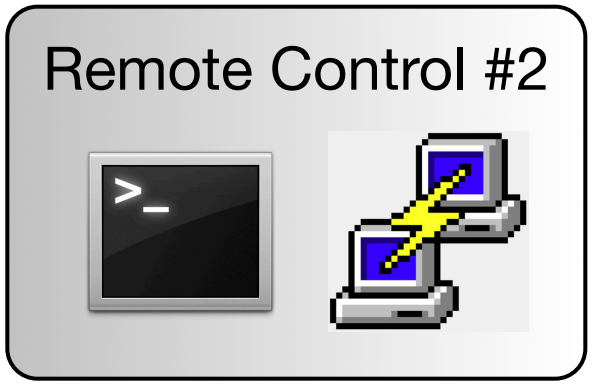
```
$ mkdir day1
```

```
$ cd day1
```

```
$ bowtie2 -x ../genomes/hg19 -U ../rawdata/exome.fastq -S exome.sam
```

```
2720452 reads; of these:
  2720452 (100.00%) were unpaired; of these:
    9467 (0.35%) aligned 0 times
    1912279 (70.29%) aligned exactly 1 time
    798706 (29.36%) aligned >1 times
99.65% overall alignment rate
```

Task 11: Monitor activity after opening a second SSH connection

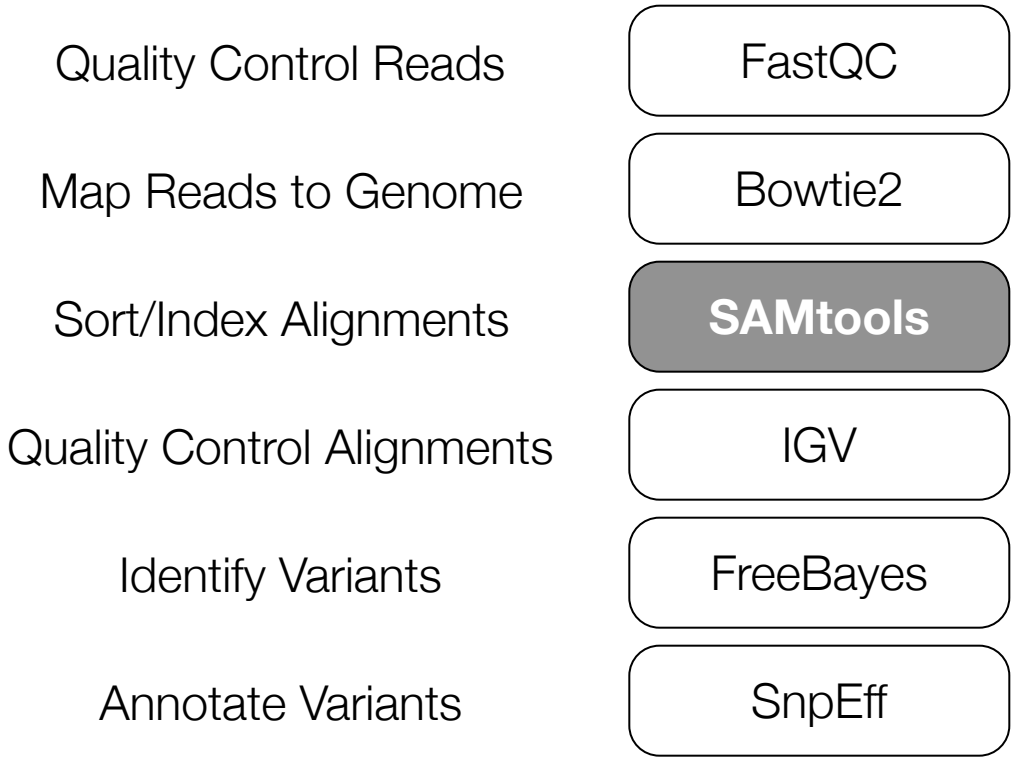



```
$ top
```

```
top - 04:49:11 up 12:37,  2 users,  load average: 0.22, 0.06, 0.06
Tasks:  70 total,   1 running,  69 sleeping,   0 stopped,   0 zombie
%Cpu(s): 99.7 us,  0.0 sy,  0.0 ni,  0.0 id,  0.3 wa,  0.0 hi,  0.0 si,  0.0 st
KiB Mem:    507228 total,   500760 used,     6468 free,     1984 buffers
KiB Swap:  5241852 total,     1924 used,  5239928 free.   372776 cached Mem

  PID USER      PR  NI    VIRT    RES    SHR S %CPU %MEM     TIME+ COMMAND
 2350 workshop  20   0  195112  86620   1916 S 99.1 17.1   0:14.52 bowtie2-ali+
    1 root      20   0    4196   1700   1072 S  0.0  0.3   0:00.51 init
    2 root      20   0       0      0      0 S  0.0  0.0   0:00.00 kthreadd
    3 root      20   0       0      0      0 S  0.0  0.0   0:00.19 ksoftirqd/0
    5 root       0 -20       0      0      0 S  0.0  0.0   0:00.00 kworker/0:0H
    6 root      20   0       0      0      0 S  0.0  0.0   0:04.89 kworker/u2:0
    7 root      20   0       0      0      0 S  0.0  0.0   0:00.65 rcu_sched
    8 root      20   0       0      0      0 S  0.0  0.0   0:00.00 rcu_bh
    9 root      rt   0       0      0      0 S  0.0  0.0   0:00.00 migration/0
   10 root      rt   0       0      0      0 S  0.0  0.0   0:01.03 watchdog/0
   11 root       0 -20       0      0      0 S  0.0  0.0   0:00.00 khelper
   12 root      20   0       0      0      0 S  0.0  0.0   0:00.00 kdevtmpfs
   13 root       0 -20       0      0      0 S  0.0  0.0   0:00.00 netns
   14 root       0 -20       0      0      0 S  0.0  0.0   0:00.00 writeback
   15 root       0 -20       0      0      0 S  0.0  0.0   0:00.00 kintegrityd
   16 root       0 -20       0      0      0 S  0.0  0.0   0:00.00 bioset
   17 root       0 -20       0      0      0 S  0.0  0.0   0:00.07 kworker/u3:0
```

Exome-seq Analysis Pipeline



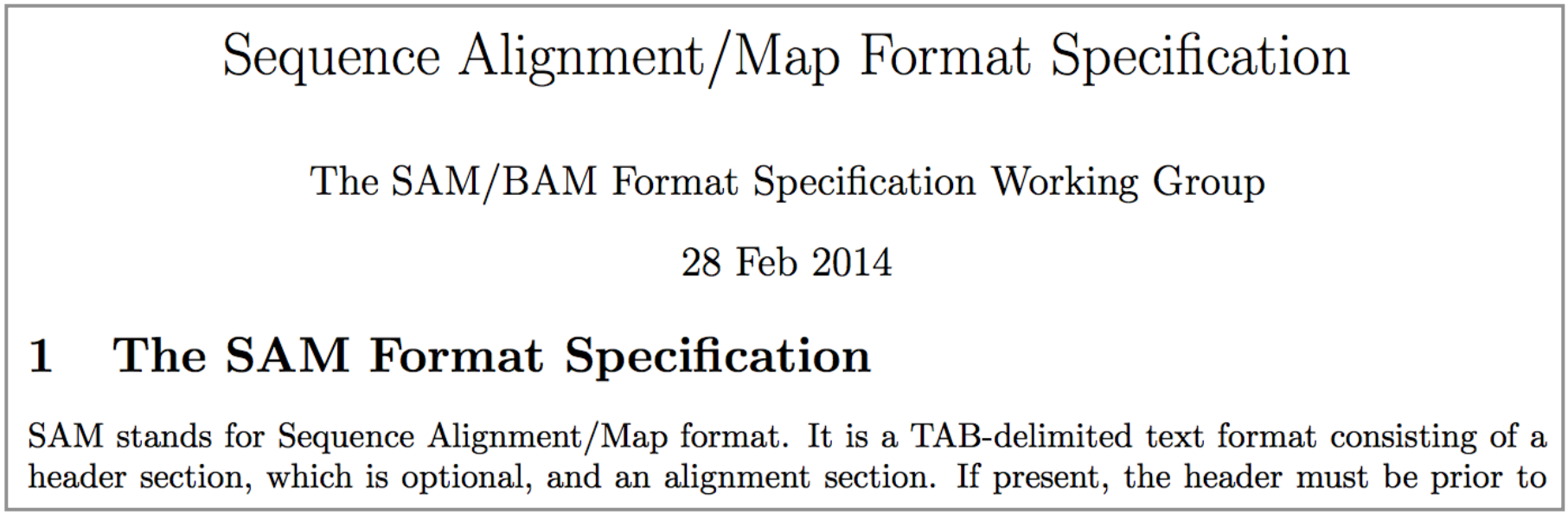
Task 12: View Bowtie2 .SAM output

```
$ ls -l
```

```
-rw-rw-r-- 1 workshop workshop 845M Aug 10 03:36 exome.sam
```

```
$ less -S exome.sam
```

```
@HD     VN:1.0  SO:unsorted
@SQ     SN:chr19        LN:59128983
@PG     ID:bowtie2      PN:bowtie2      VN:2.1.0
SRR925785.2     0       chr19   59118851        0       17M1D84M        *       0       0       GGTT
SRR925785.3     16      chr19   246360  0       16M4I81M        *       0       0       CCTCGCGGTACC
SRR925785.327   0       chr19   60018   42      101M    *       0       0       TGCTCCCCACCCTCTGCACA
SRR925785.328   0       chr19   60307   0       69M3I3M1I25M    *       0       0       ATGGGGCAAGCA
SRR925785.329   0       chr19   60150   42      101M    *       0       0       AGGATAATCTGACCTGCAGG
SRR925785.330   0       chr19   60970   42      101M    *       0       0       ATGGTGGAGGGGTTTGAGAA
SRR925785.331   0       chr19   60951   42      101M    *       0       0       GGCAGGAGGAGGGTGTGGGA
```



Sequence Alignment/Map Format Specification

The SAM/BAM Format Specification Working Group

28 Feb 2014

1 The SAM Format Specification

SAM stands for Sequence Alignment/Map format. It is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section. If present, the header must be prior to

samtools.sourceforge.net/SAM1.pdf

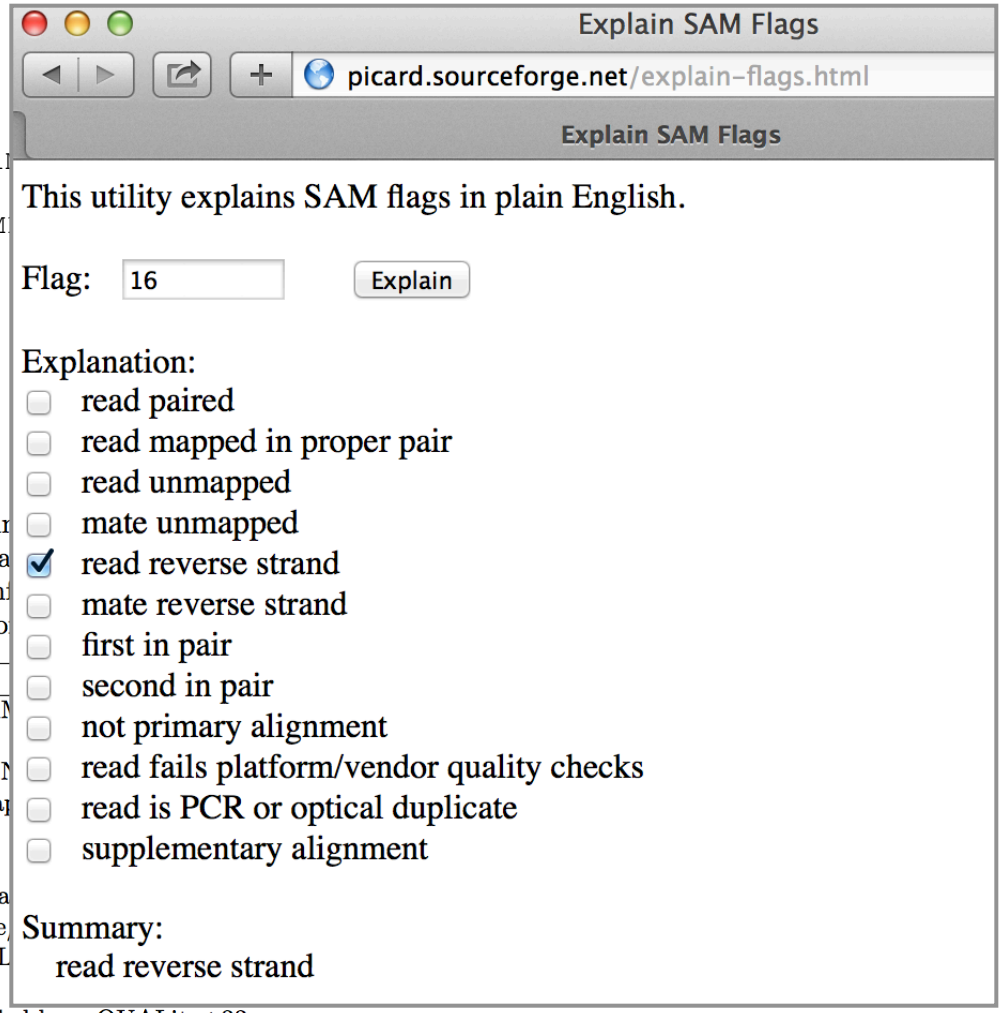
SAM Flags Explained

```
@HD     VN:1.0  SO:unsorted
@SQ     SN:chr19        LN:59128983
@PG     ID:bowtie2      PN:bowtie2      VN:2.1.0
SRR925785.2     0       chr19   59118851        0
SRR925785.3     16      chr19   246360  0       16M4I81
SRR925785.327   0       chr19   60018   42      101M
SRR925785.328   0       chr19   60307   0       69M3I3M
SRR925785.329   0       chr19   60150   42      101M
SRR925785.330   0       chr19   60970   42      101M
SRR925785.331   0       chr19   60951   42      101M
```

1.4 The alignment section: mandatory fields

In the SAM format, each alignment line typically represents the linear alignr
line has 11 mandatory fields. These fields always appear in the same order a
their values can be '0' or '*' (depending on the field) if the corresponding in
The following table gives an overview of the mandatory fields in the SAM fo

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	`[!-?A-~]{1,255}`	Query template NAM
2	FLAG	Int	$[0,2^{16}-1]$	bitwise FLAG
3	RNAME	String	`\*|[!-()+-<>-~][!-~]*`	Reference sequence N
4	POS	Int	$[0,2^{31}-1]$	1-based leftmost map
5	MAPQ	Int	$[0,2^{8}-1]$	MAPping Quality
6	CIGAR	String	`\*|([0-9]+[MIDNSHPX=])+`	CIGAR string
7	RNEXT	String	`\*|=|[!-()+-<>-~][!-~]*`	Ref. name of the ma
8	PNEXT	Int	$[0,2^{31}-1]$	Position of the mate
9	TLEN	Int	$[-2^{31}+1,2^{31}-1]$	observed Template L
10	SEQ	String	`\*|[A-Za-z=.]+`	segment SEQuence
11	QUAL	String	`[!-~]+`	ASCII of Phred-scaled base QUALity+33



Task 13: Summarize alignments using `cut`, `sort`, and `uniq`

```
$ cut -f2 exome.sam
```

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

```
$ cut -f2 exome.sam | head
```

```
$ cut -f2 exome.sam | head | sort
```

```
$ cut -f2 exome.sam | head | sort | uniq
```

```
$ cut -f2 exome.sam | head | sort | uniq -c
```

```
$ cut -f2 exome.sam | sort | uniq -c
```

```
1375841 0
1335144 16
   9467 4
      1 ID:bowtie2
      1 SN:chr19
      1 VN:1.0
```

SAMtools Conversion Suite

```
$ samtools
```

```
Program: samtools (Tools for alignments in the SAM format)
Version: 0.1.19-44428cd

Usage:   samtools <command> [options]

Command: view         SAM<->BAM conversion
         sort         sort alignment file
         index        index alignment
         idxstats     BAM index stats (r595 or later)
         merge        merge sorted alignments
         rmdup        remove PCR duplicates
         bedcov       read depth per BED region
```

```
$ samtools view
```

```
Usage:   samtools view [options] <in.bam>|<in.sam> [region1 [...]]

Options: -b       output BAM
         -h       print header for the SAM output
         -H       print header only (no alignments)
         -S       input is SAM
                                ...
         -f INT   required flag, 0 for unset [0]
         -F INT   filtering flag, 0 for unset [0]
```

Task 14: Summarize alignments using `samtools view`

<UP>

```
$ samtools view -S -f16 exome.sam
```

<CTRL>-C

<UP>

```
$ samtools view -S -f16 exome.sam | wc -l
```

```
1335144
```

Task 15: Convert, sort, and index .BAM file using SAMtools

```
$ samtools view -bS exome.sam > exome.bam
```

```
$ ls
```

```
exome.sam   exome.bam
```

```
$ samtools sort -m 400000000 exome.bam exome.sorted
```

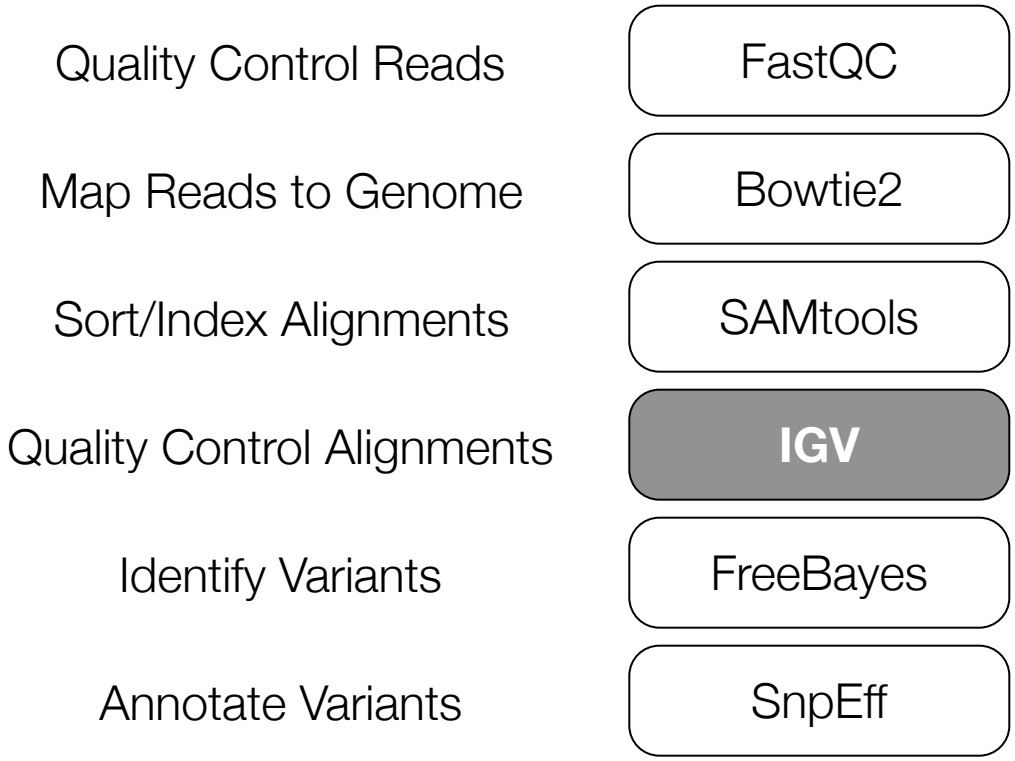
```
              or -m 400M
```

```
$ samtools index exome.sorted.bam
```

```
$ ls
```

```
exome.sam   exome.bam   exome.sorted.bam   exome.sorted.bam.bai
```

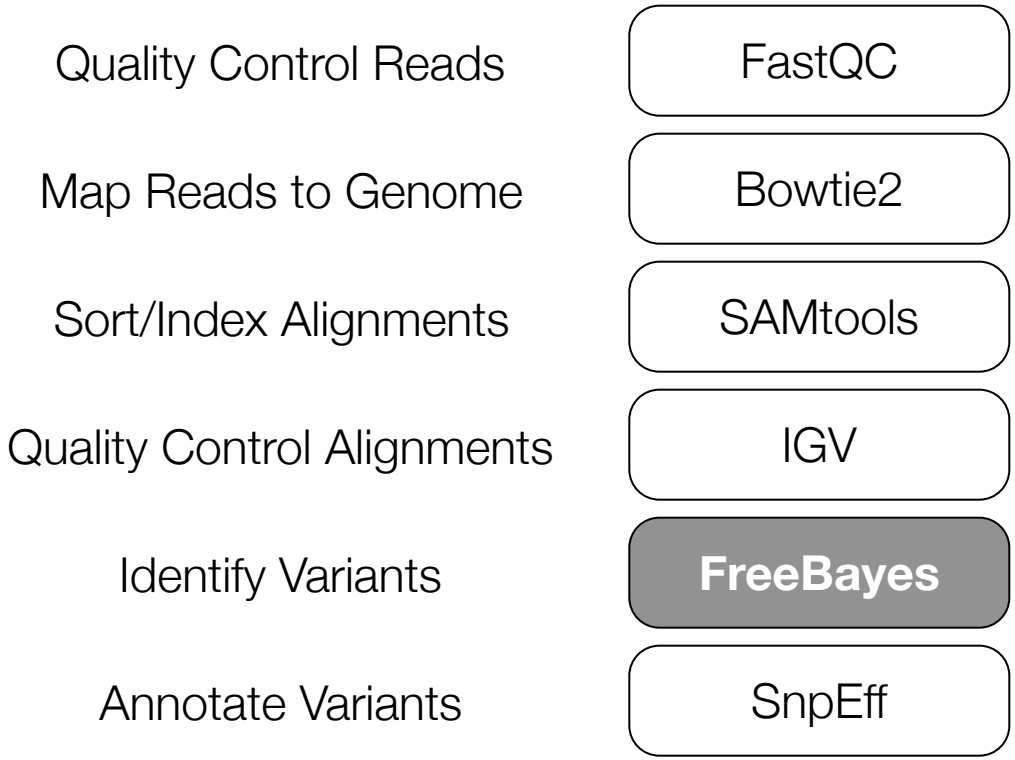
Exome-seq Analysis Pipeline



Task 16: Visualize genomics data with Integrative Genomics Viewer

broadinstitute.org/igv

Exome-seq Analysis Pipeline



FreeBayes Variant Detection

```
$ freebayes
```

```
usage: freebayes -f [REFERENCE] [OPTIONS] [BAM FILES] >[OUTPUT]

Bayesian haplotype-based polymorphism discovery.

parameters:

   -h --help       For a complete description of options.
```

By default, FreeBayes will consider variants supported by **at least 2 observations** in a single sample (-C) and also by **at least 20% of the reads** from a single sample (-F).

Ploidy may be set to any level (-p), but by default all samples are assumed to be diploid. FreeBayes **can model per-sample and per-region variation in copy-number (-A)** using a copy-number variation map.

Task 17: Start identifying putative variants using FreeBayes

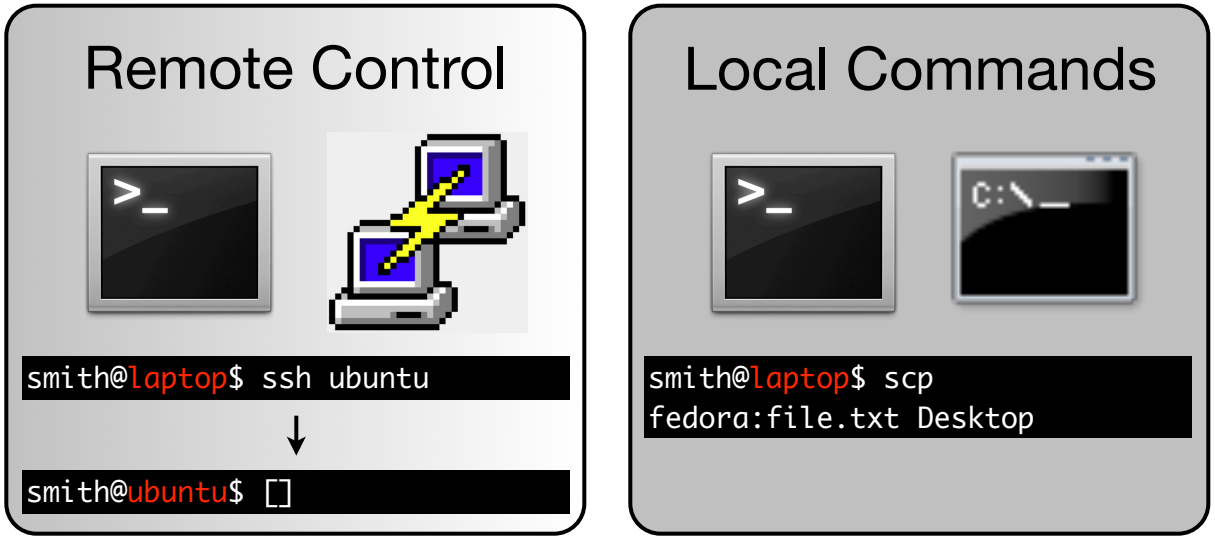
```
$ freebayes -f ../genomes/hg19.fa --pooled-discrete
  --pooled-continuous --min-alternate-fraction 0.05
  --genotype-qualities --report-genotype-likelihood-max
  exome.sorted.bam > exome.vcf
```

<SARAH EXPLAINS FREEBAYES MAGIC WHILE WE WAIT>

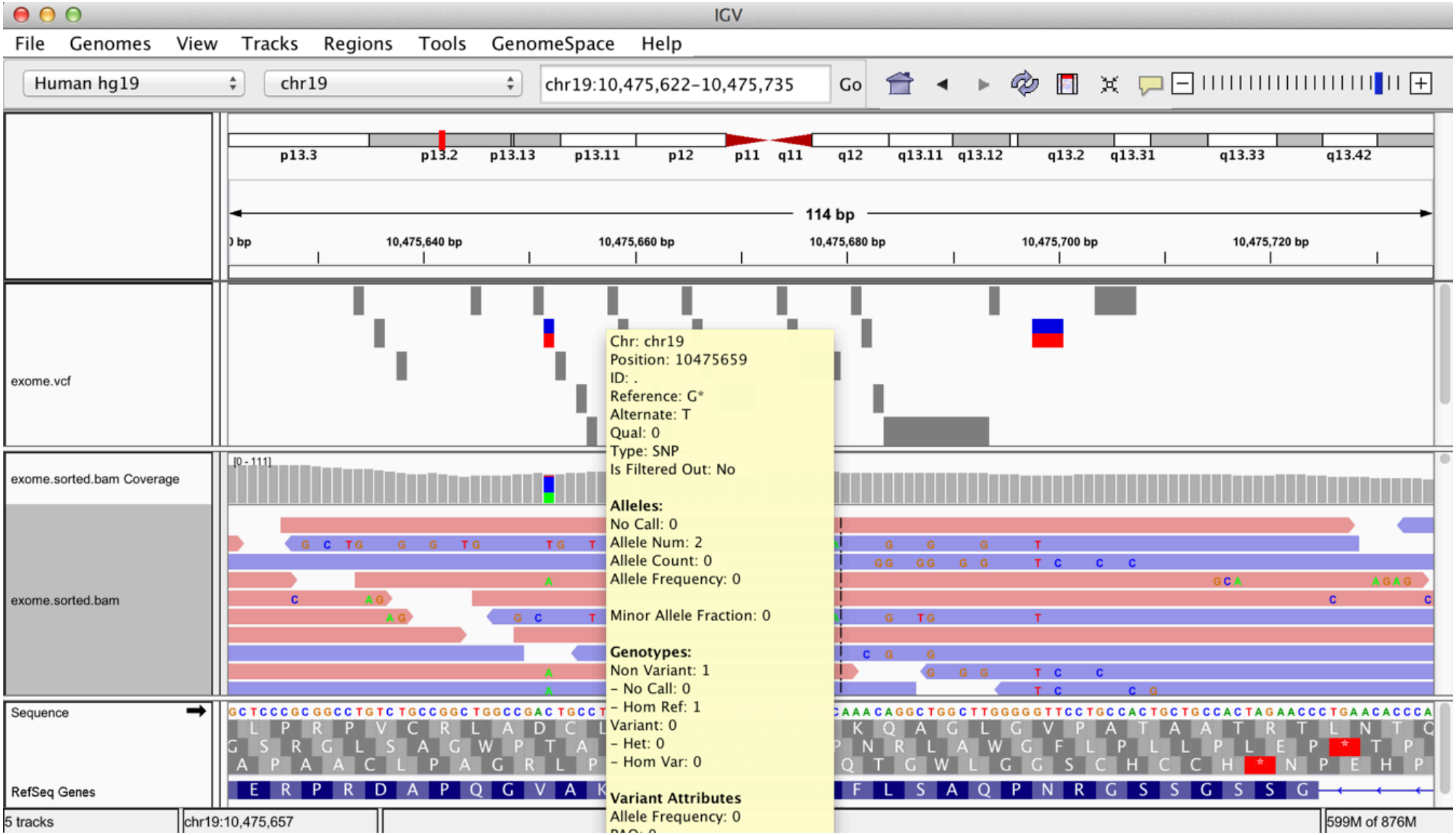
Task 18: Transfer `exome.vcf` to your laptop using `scp`

```
$ cd Desktop/PG2014
```

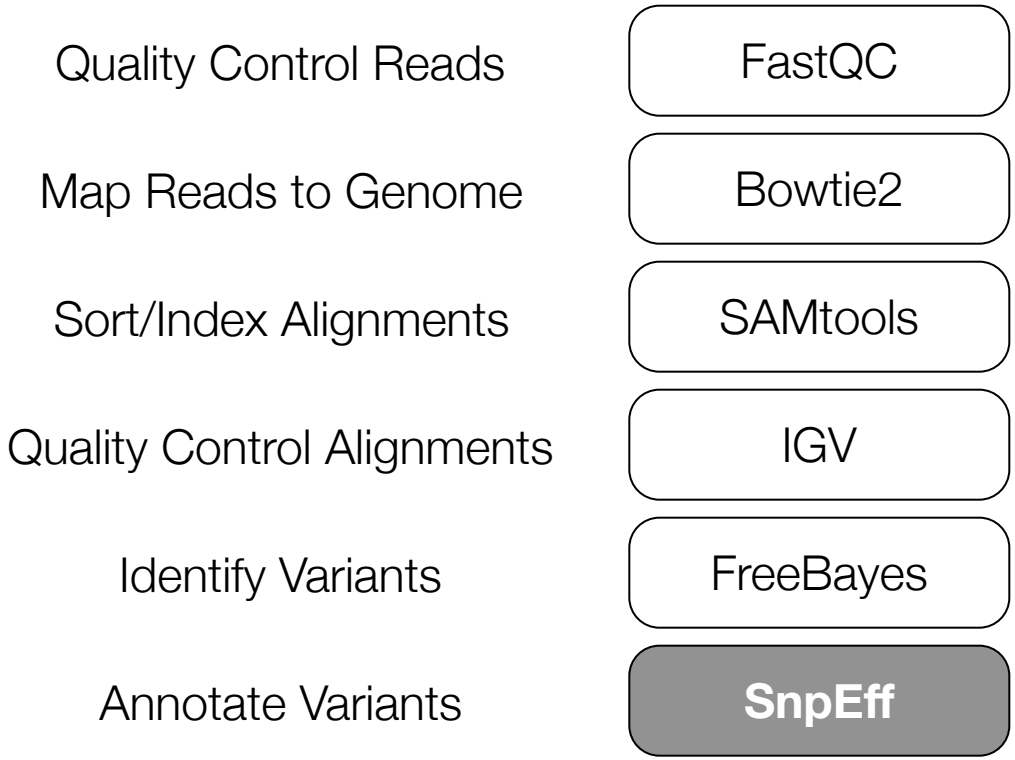
```
$ scp workshop@192.168.56.103:day1/exome.vcf .
```



Task 19: **Overlay** `exome.sorted.bam` **and** `exome.vcf` **in IGV**



Exome-seq Analysis Pipeline



Task 20: View contents of variant calls

```
$ less -S exome.vcf
```

```
##fileformat=VCFv4.1
##fileDate=20140804
##source=freeBayes v0.9.14-25-g37293e3
##reference=../genomes/hg19.fa
##phasing=none
##commandline="freebayes -f ../genomes/hg19.fa --pooled-discrete --pooled-contin
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of samples with data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total read depth at the locus">
##INFO=<ID=DPB,Number=1,Type=Float,Description="Total read depth per bp at the l
```

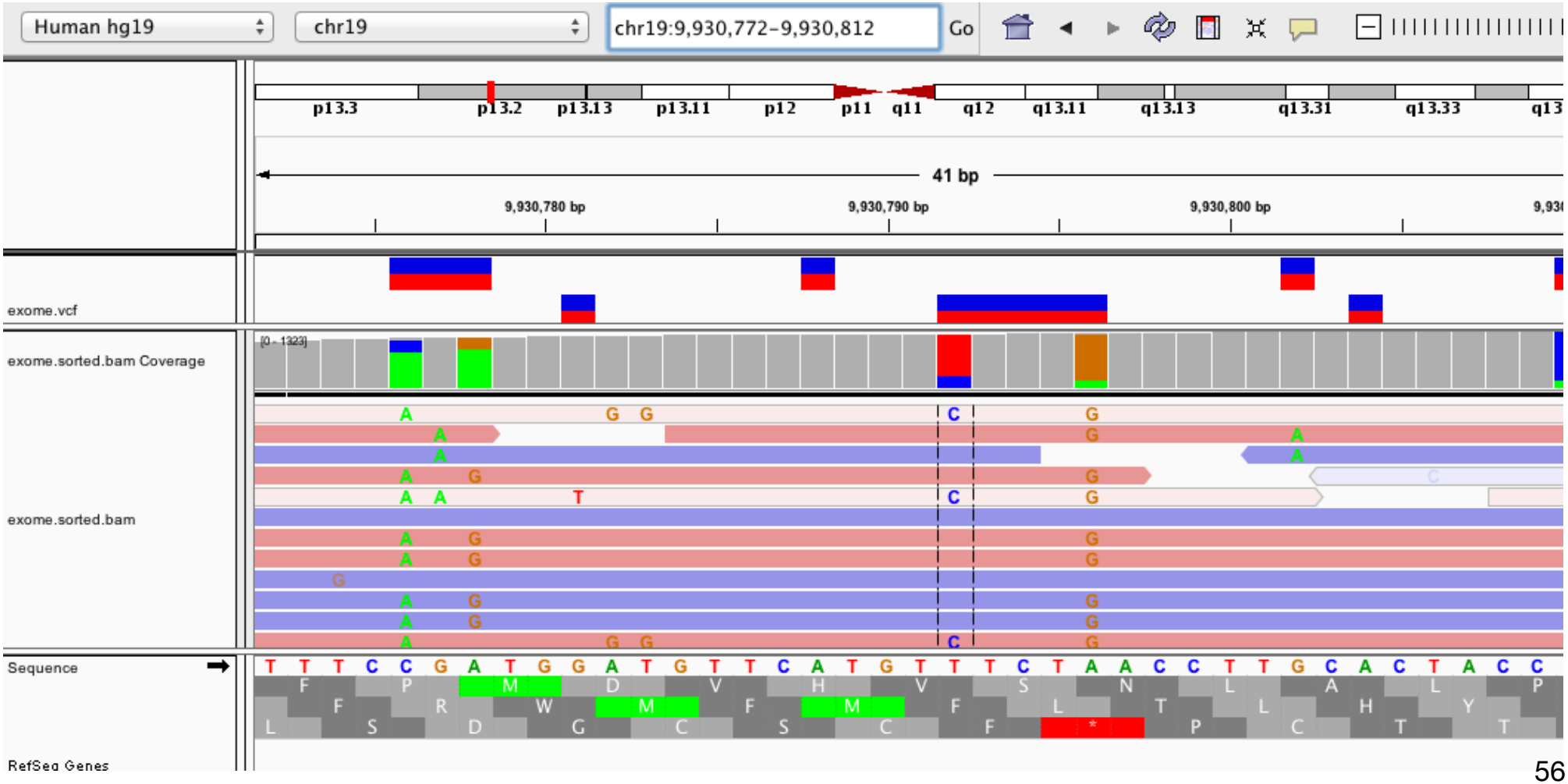
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	unknown
chr19	60030	.	T	C	0.000229488		.	AB=0;ABP=0;AC=0;	
chr19	60037	.	A	C	8.95987e-05		.	AB=0;ABP=0;AC=0;	
chr19	60107	.	A	G	44.3113	.	AB=0.428571;ABP=3.32051;		
chr19	60319	.	C	T	55.5081	.	AB=0.363636;ABP=4.78696;		
chr19	60354	.	T	A	167.575	.	AB=0.538462;ABP=3.17734;		
chr19	60408	.	A	C	1.25744e-05		.	AB=0;ABP=0;AC=0;	
chr19	60454	.	A	C	51.3931	.	AB=0.5;ABP=3.0103;AC=1;A		
chr19	60479	.	CCC	TCT	97.3686	.	AB=0.833333;ABP=8.80089;		

1000genomes.org/wiki/Analysis/Variant+Call+Format/vcf-variant-call-format-version-41

Task 21: Sort variant calls by QUAL score

```
$ sort -k6nr exome.vcf | less -S
```

```
... 157026 highest score
... 9930792 haplotype example
```



AWK Splits Lines into Separate Fields

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
chr19	60030	.	T	C	0.000229488	.	**AB=0;ABP=0**
chr19	60037	.	A	C	8.95987e-05	.	**AB=0;ABP=0**
chr19	60107	.	A	G	44.3113	.	AB=0.42857
chr19	60319	.	C	T	55.5081	.	AB=0.36363
chr19	60354	.	T	A	167.575	.	AB=0.53846
chr19	60408	.	A	C	1.25744e-05	.	**AB=0;ABP=0**
chr19	60454	.	A	C	51.3931	.	**AB=0.5;ABP**
chr19	60479	.	CCC	TCT	97.3686	.	AB=0.83333
$1	**$2**	**$3**	**$4**	**$5**	**$6**	**$7**	**$8**

Task 22: Filter and manipulate using AWK

```
$ awk '$6 > 20' exome.vcf | wc -l
```

```
...  59,545
```

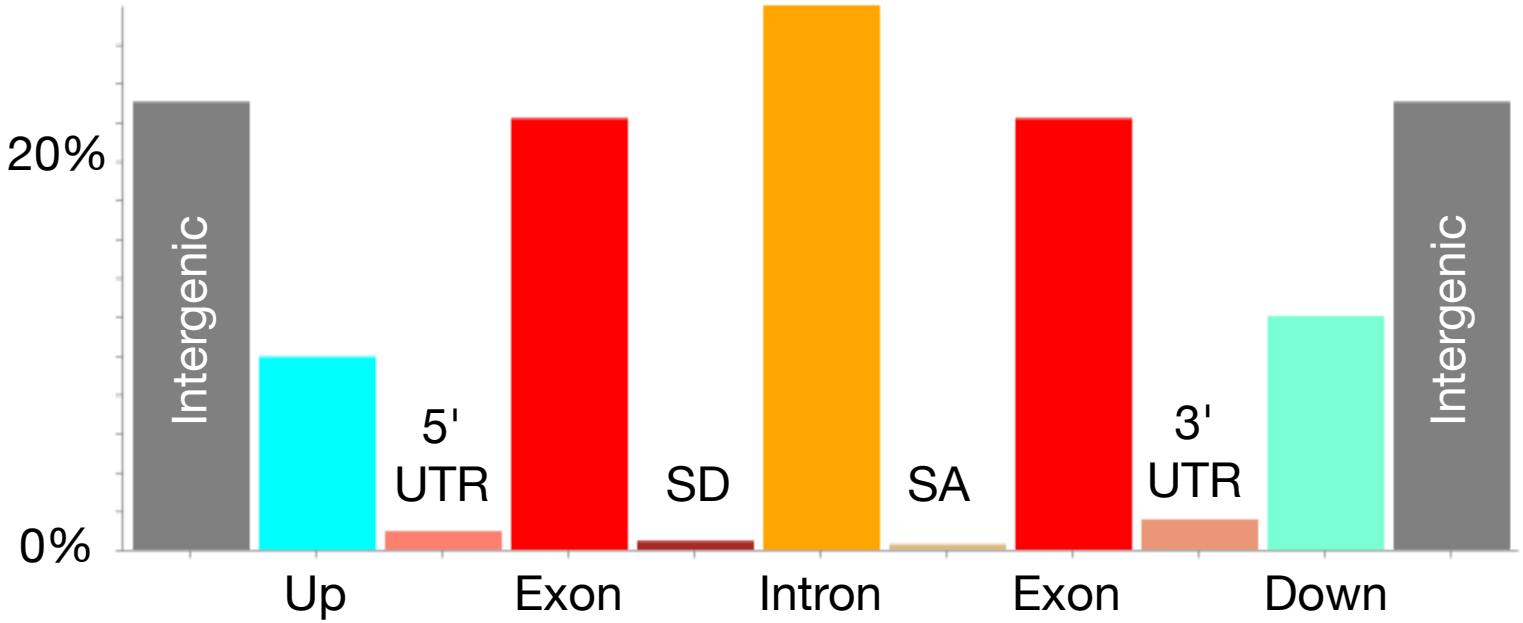
```
$ wc -l exome.vcf
```

```
... 199,169
```

```
$ tail exome.vcf | awk '{print $4 $2 $5}'
```

```
GTT59118889GT
T59118892C
A59118893G
```

SnpEff Annotates and Prioritizes Variants



Gene(s) Affected

EXON_DELETED, STOP_GAINED, FRAME_SHIFT
CODON_CHANGE, MICRO_RNA, INTRAGENIC

ENCODE, Epigenome Roadmap, NextProt, ENSEMBL/Jaspar motifs

Task 23: Annotate variants with SnpEff

```
$ cp ~/answers/day1/snpEff.config .
```

```
$ java -jar ~/src/snpEff/snpEff.jar eff hg19 exome.vcf >
  exome.annotated.vcf
```

Summary

Genome	hg19
Date	2014-08-07 11:48
SnpEff version	`SnpEff 3.6c (build 2014-05-20), by Pablo Cingolani`
Command line arguments	`SnpEff  hg19 exome.vcf`
Warnings	70,091
Number of lines (input file)	199,112
Number of variants (before filter)	215,845
Filter	
Number of variants filtered out	0
Number of not variants (i.e. reference equals alternative)	0
Number of variants processed (i.e. after filter and non-variants)	215,845
Number of known variants (i.e. non-empty ID)	0 (0%)
Number of effects	290,606
Genome total length	59,128,983
Genome effective length	59,128,983
Change rate	1 change every 273 bases

Task 24: Explore `snpEff_genes.txt`

```
$ cut -f1,43,54 snpEff_genes.txt | head
```

```
# The following table is formatted as tab separated values.
#GeneId Count (FRAME_SHIFT) Count (STOP_GAINED)
A1BG  0 0
A1BG-AS1  0 1
ABCA7 0 6
ABHD17A 2 7
```

```
$ awk '{total=$43+$54; print $1, $43, $54, total}'
snpEff_genes.txt | head
```

```
$ awk '{total=$43+$54; print $1, $43, $54, total}'
snpEff_genes.txt | sort -k4nr | head
```

```
GMIP 0 15 15
ZNF614 2 13 15
CYP2A7 0 13 13
FCGBP 0 13 13
ZNF527 0 12 12
```

Task 25: Filter variants based on effect

```
$ grep STOP_GAINED exome.annotated.vcf | wc -l
```

```
... 611
```

heterozygous vs homozygous?

```
$ grep STOP_GAINED exome.annotated.vcf | grep ";AF=1;" | wc -l
```

```
... 6
```

non-essential in cell culture?

```
$ grep STOP_GAINED exome.annotated.vcf | grep ";AF=1;" | less -S
```

```
$ grep FRAME_SHIFT exome.annotated.vcf | grep ";AF=1;" | wc -l
```

```
... 5
```

52803669

```
$ grep FRAME_SHIFT exome.annotated.vcf | grep ";AF=1;" | less -S
```

Exome-seq Analysis Pipeline

Quality Control Reads	FastQC	v0.11.2
Map Reads to Genome	Bowtie2	v2.1.0
Sort/Index Alignments	SAMtools	v0.1.19
Quality Control Alignments	IGV	v2.3.34
Identify Variants	FreeBayes	v0.9.14
Annotate Variants	SnpEff	v3.6c

hg19 (w/o alternate haplotypes)

UCSC knownGene Aug 2014

Methods Section / SRA Metadata

MCF10F (*H.* sapiens, breast cancer, non-malignant)

Qiagen Gentra Puregene Cell Kit
Covaris E220 to 150 bp
Agilent Automation Systems w/ 8 cycles of PCR
Agilent SureSelect XT 37Mb Capture Library w/ 12 cycles of PCR
Illumina Genome Analyzer IIx for 101 cycles

2,720,452 reads were aligned to hg19 with Bowtie2 v2.1.0 using default parameters
(70.29% aligned exactly one time, 29.36% aligned >1 times)

Variants were identified with FreeBayes v0.9.14 using the following parameters

```
freebayes -f hg19.fa --pooled-discrete --pooled-continuous
          --min-alternate-fraction 0.05 --genotype-qualities
          --report-genotype-likelihood-max
```