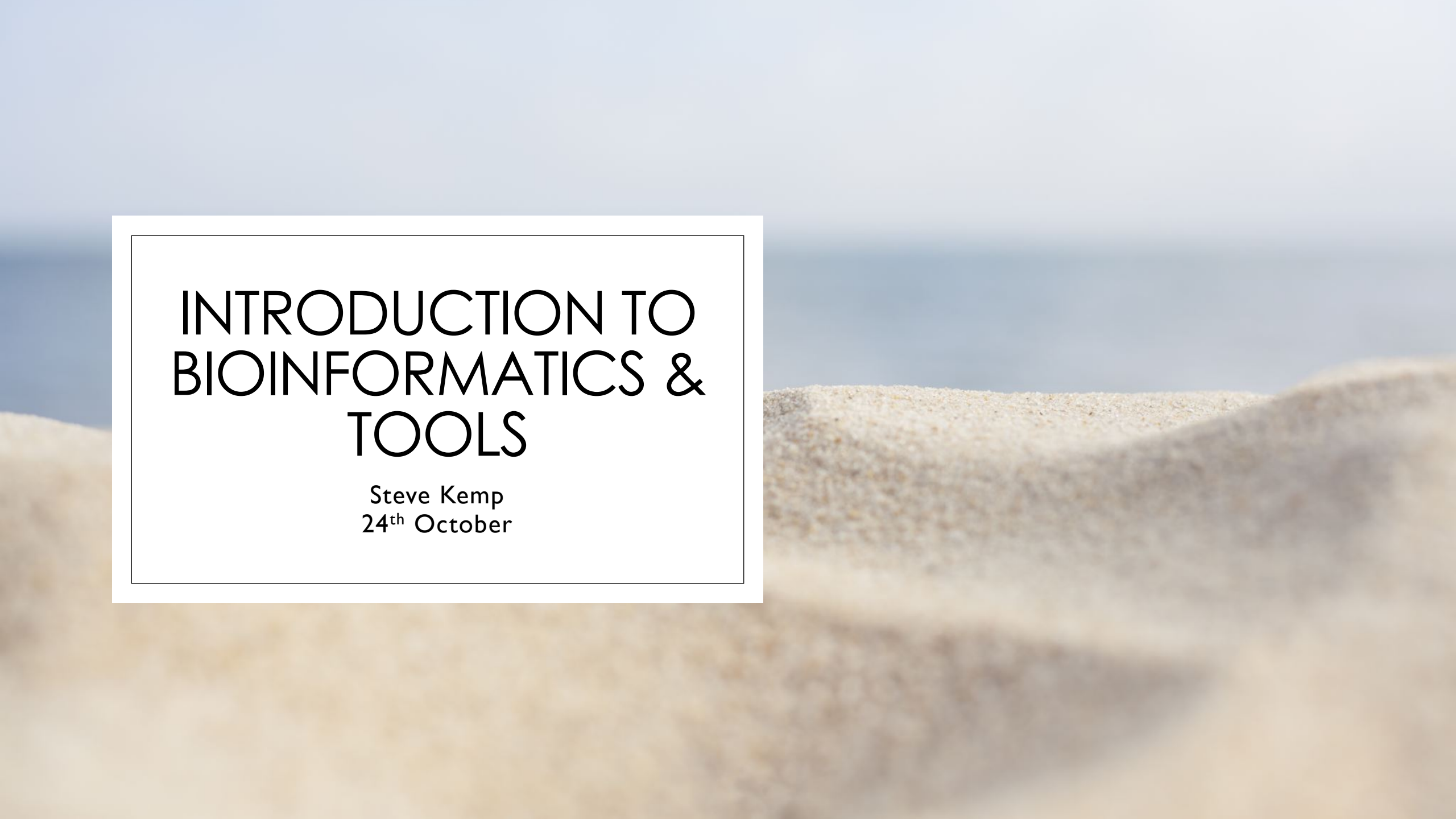
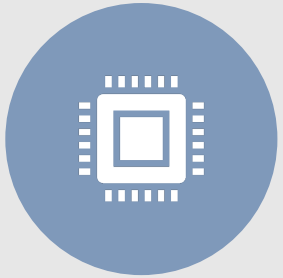


INTRODUCTION TO BIOINFORMATICS & TOOLS

Steve Kemp

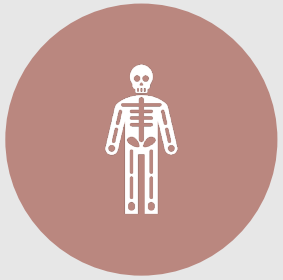
24th October

What is Bioinformatics?



An **interdisciplinary field** that uses **software** to understand **biological data**.

Combination of **biology, computer science, information engineering, mathematics** and **statistics**.

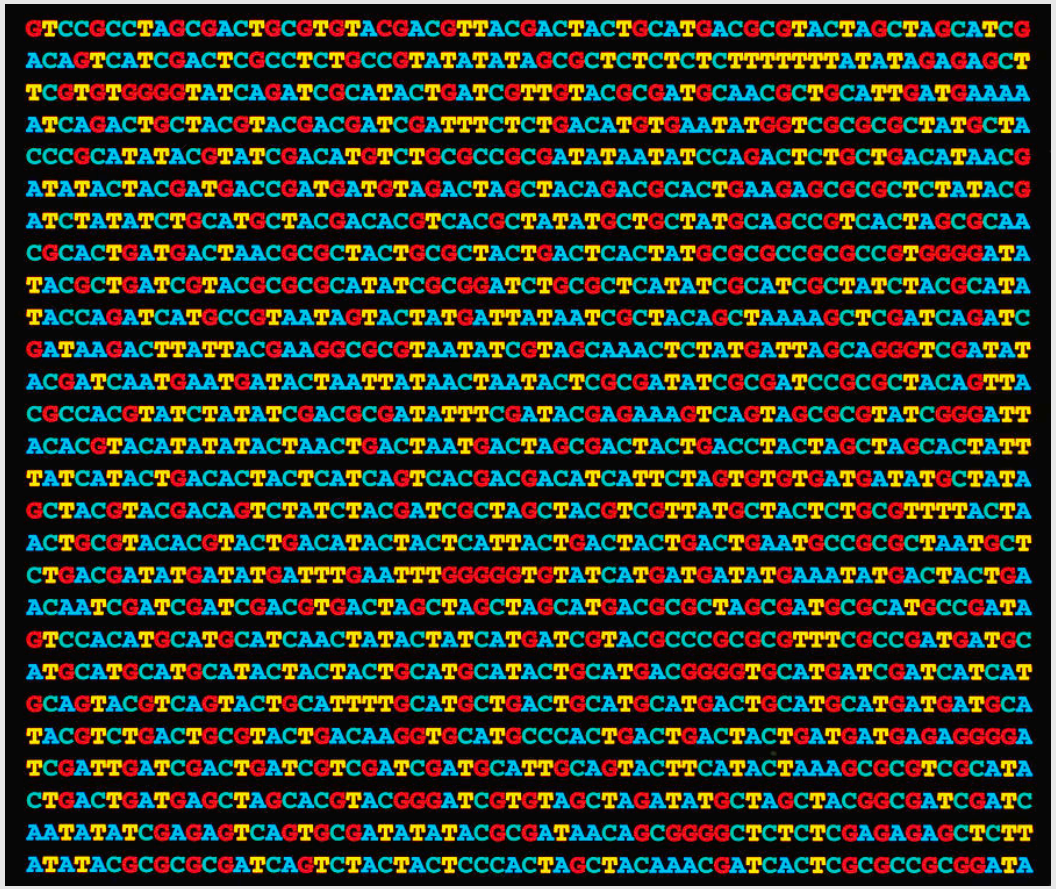


Bioinformatics tools can be used individually to answer a specific question or chained together into a "**pipeline**" for rapid analyses.

For **genetic analyses**, bioinformatics analyses **sequence data** and allows us to **annotate genomes**, determine **mutations** and reconstruct predicted **ancestral phylogenies.**

Sequence Data

- A "sequence" is the order of **nucleotides** or **amino acids** that make up a strand of **DNA**.
- Many common sequencing methods produce **short fragments** of DNA called "**reads**".
- Raw reads must be **processed** and **assembled** to obtain a **complete genome**.
- Raw reads are comprised of **forward** and **reverse** reads.



Basic Bioinformatics Pipeline

Download raw forward + reverse reads (.fastq)

Pre-alignment quality control (FastQC)

Mapping reads to a reference (bwa)

Sort & Index BAM files (samtools)

Eyeballing the alignment (IGV)

Post-alignment quality control (FastQC)

Downstream analysis

Bioinformatic Tools

- Analysis of whole genomes can be **computationally expensive**.
- Many tools are run on Linux cluster computers which have faster **CPUs** with more **threads**.
- Most common (free) bioinformatic tools are used by typing **code** into a **terminal** – it is best to use **Linux** or a **Mac** for Bioinformatics (both are based on UNIX). If using a Windows PC/Laptop, consider installing **Ubuntu 18.04.**
- Installing various software on a Mac or Linux system can be managed using **Homebrew** or **Conda.**

If you already installed Ubuntu according to the instructions sent out last week, then feel free to click along with me.

Mac vs Linux

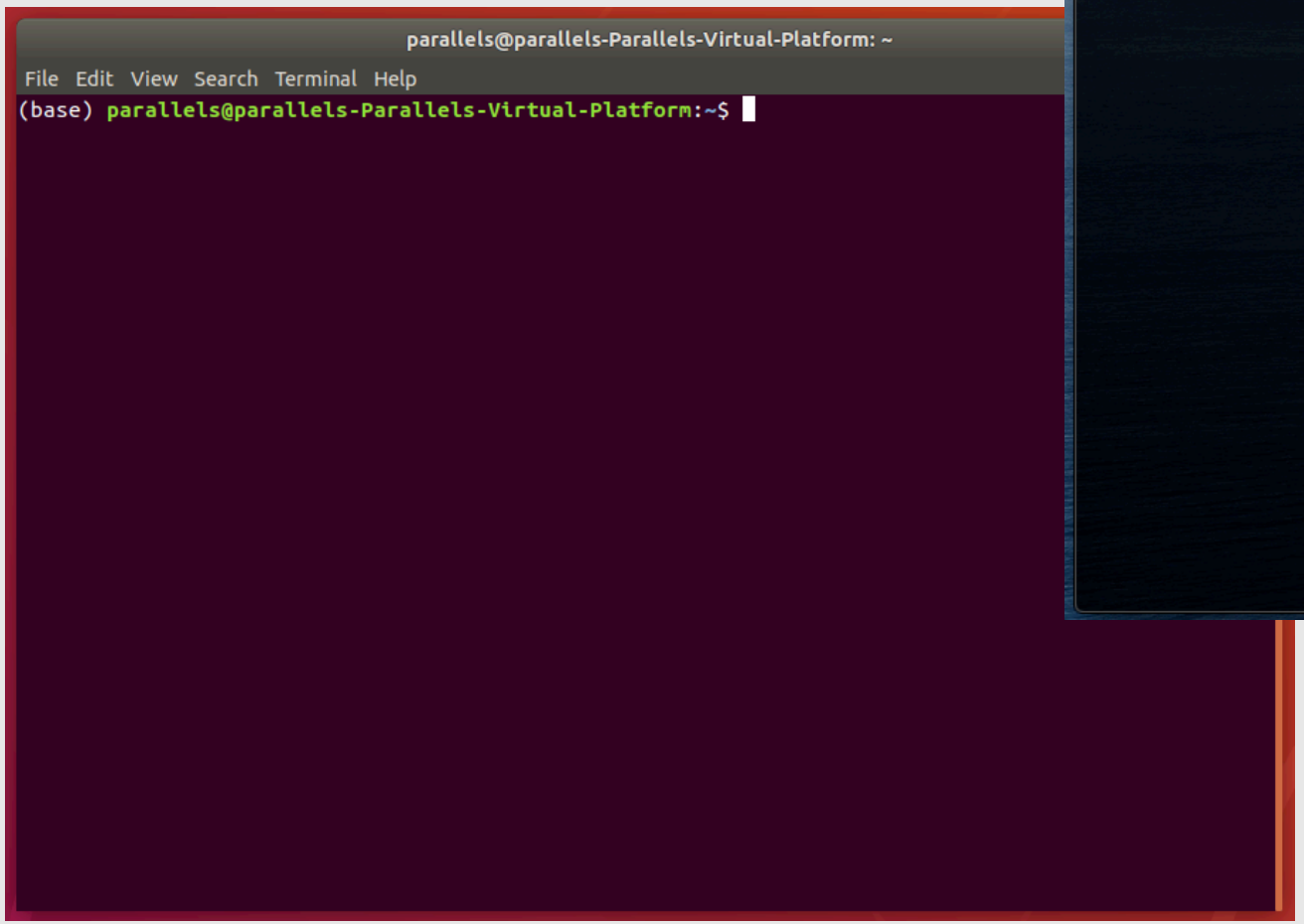
- This webinar is primarily geared towards **Linux users**.
- Installation of all software is tailored to **Linux machines** – installation inputs (anything with sudo in front of it) **will not work** on a mac.
- Once you have installed the software, then the **usage commands** are identical on Linux or Macs.
- If you need help installing any software, please email me or have a google.
- If using a mac, almost all of the discussed software can be installed via homebrew or conda, or by downloading the .dmg files from the software websites on a mac ☺

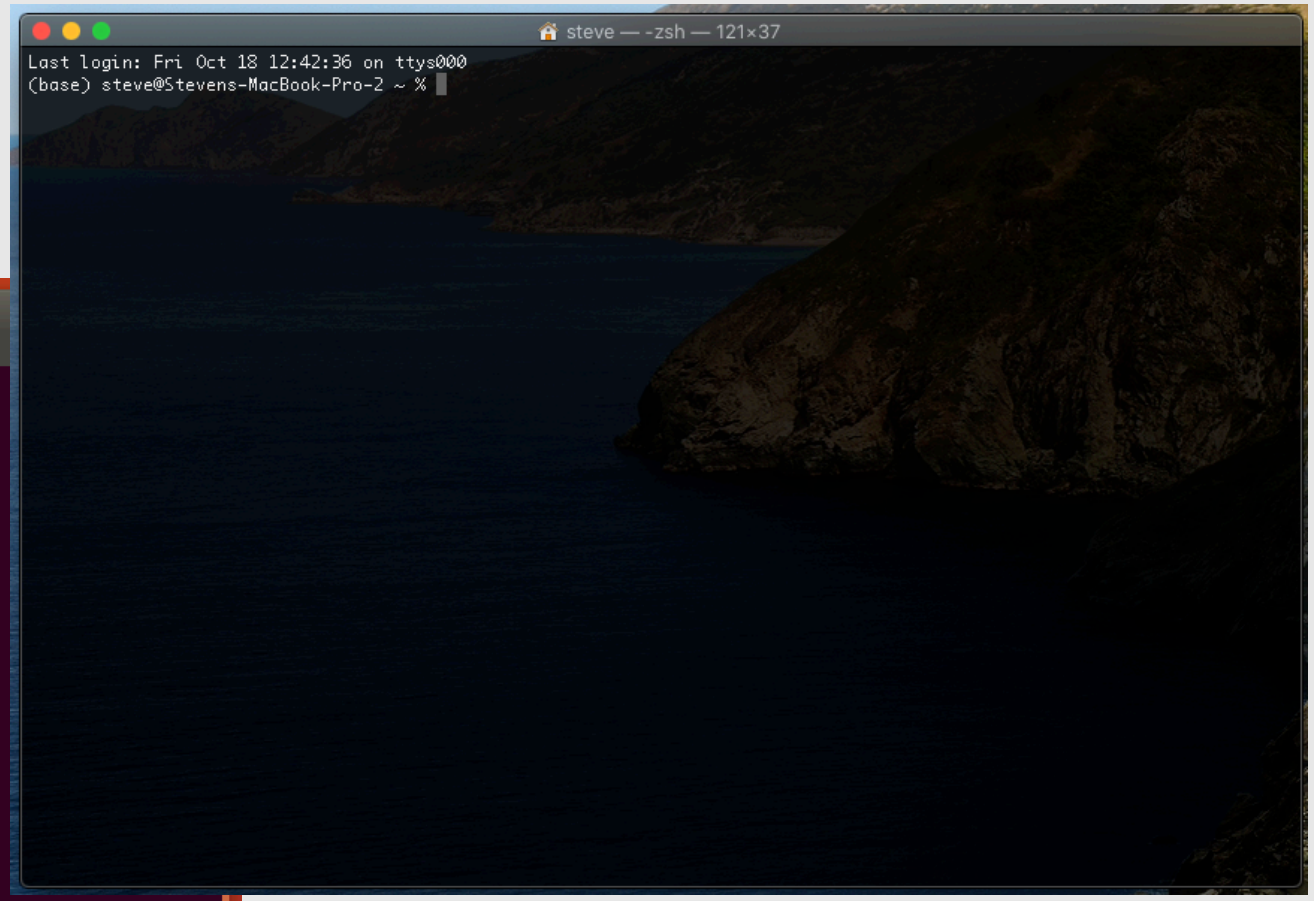
Terminals


Activities
Fri 13:36
Type to search...
Mahjongg
Mines
Power Statistics
R
Remmina
Rhythmbox
Settings
Shotwell
Simple Scan
Software & Up...
Software Upd...
Startup Applic...
Startup Disk C...
Sudoku
System Monitor
Terminal
Frequent
All


Other
QuickTime Player
TextEdit
Grapher
Time Machine
Font Book
Chess
Stickies
Image Capture
VoiceOver Utility
AirPort Utility
Migration Assistant
Terminal
Activity Monitor
Console
Keychain Access
System Information
Automator
Script Editor
Disk Utility
Boot Camp Assistant
Digital Colour Meter
ColorSync Utility
Screenshot
Bluetooth File Exchange
Audio MIDI Setup
XQuartz
Feedback Assistant

Terminals


parallels@parallels-Parallels-Virtual-Platform: ~
File Edit View Search Terminal Help
(base) parallels@parallels-Parallels-Virtual-Platform:~$


steve — -zsh — 121×37
Last login: Fri Oct 18 12:42:36 on ttys000
(base) steve@Stevens-MacBook-Pro-2 ~ %

Basic Linux/Unix commands

- To navigate the terminal you type code, rather than point and click.
- The directory that you are in when you first open the terminal is the HOME directory.
- To view the contents of the HOME directory, type **ls** and hit enter.
- Most of this tutorial will take place in the DOWNLOADS directory. To change directory, type **cd Downloads/**

- **ls = list directory contents**
- **cd = change directory // cd ../ go up one directory**
- **mkdir {name} = make directory**
- **mv = change filename or move from one folder to another**
- **rm = remove file // rm –r = remove directory**

Installing Homebrew

- **Homebrew** is a large repository of tools which can be searched and will install software with a very simple command.
- **To install**
  - Open **Terminal**
  - Type **sudo apt install linuxbrew-wrapper**
  - Follow instructions (usually type Y a few times)
  - Add Linuxbrew to your PATH:

    export PATH="/home/linuxbrew/.linuxbrew/bin:$PATH"' >>~/.bash_profile

    export MANPATH="/home/linuxbrew/.linuxbrew/share/man:$MANPATH"' >>~/.bash_profile

    echo 'export PATH="/home/linuxbrew/.linuxbrew/bin:$PATH"' >>~/.bash_profile
  - When you install software, adding to your PATH means you can call the software **anywhere** in the terminal.

Installing "Conda"

- **Miniconda2 and Minoconda3** are alternative managers similar to Homebrew.
- **To Install**:
  - Open terminal:
  - **wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh**
  - Type **bash Miniconda3-latest-Linux-x86_64.sh**
  - Follow prompts on screen (usually accept user agreement by typing yes and then choose installation path)
  - **IMPORTANT:** After installing, **close and re-open** the terminal.

Installing a Java Environment

- **Java** is a popular programming language and several common bioinformatics tools are written in this language. You need to **install a java environment**.

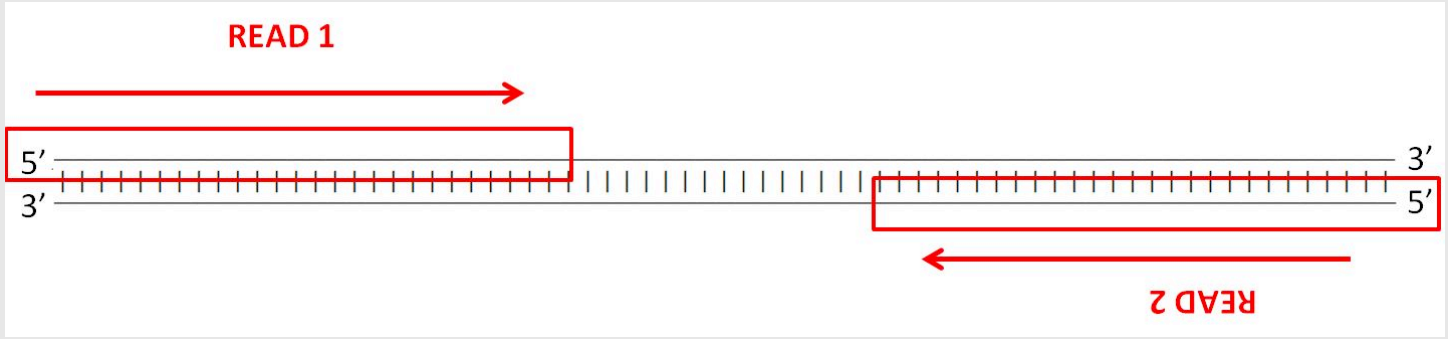
- Very easy to do, type:
  - **sudo apt install default-jre**

- You can check to see if this installed correctly by typing:
  - **java –version**

1. Download Raw reads

- A **forward** and **reverse** read is created as part of the **amplification** process during sequencing.
- At the end of the sequencing, each DNA sample will have a **raw forward** and **raw reverse** read in a file.

e.g. 14683_1_20_**1**.fastq.gz & 14683_1_20_**2**.fastq.gz



Fastq files

- Text file containing a sequence and quality scores. Quality score is coded in ASCII characters.

@MS8_14683:1:1101:1856:12754#20/1
CCCTTTCTAAGGTATTATTCCAGACCCTTCTAGTAATGTTACAATGTGCTCGTCTTATGTCTCCTAT
TATGTTCTGTGCATAGAATACCTGTCCTGGTCCTATCCTTACACTTTTTCTTGTATTATTGTTGGGT
CTTGCACACGTGATTTCTACAGATTCATTAAGATGTACTATTATTGTCTTAGCATTGTTTGATATATT
TTCAGATCTAATTATTATATCTTTTTCTGCTAGGCTACCATTTAACAGTAGTTGAGTTGACACCACT
GGCTTAATCCCATGTGTACATTGTACTATGC
+
@@@CCCEF9-
,,C,CE9ECCE9<,C@@CC<E,,C,,C,CC<C<,C,C,6CC7@@ED9C,C@CC@F,CC,C<CECE,C,<,E9,,,<,CC
C,CECF<,:CCF9CCEEFCE9CEEFEFDEEEE9EE9FF9EE8,+B=EE,,,:?A7A8,?ED?E9A<<,CAA9EF9,;,C<A;AF;A
E,EE9EAFF99,@=@EEEFF9>ADADEGGC;D,@=F9=DDADD:FCDEGFFGGGGF8EA,66A??DDD?FCD==
+C?9=F7;*BF77;=:5??FF*69AFCFCEA@A6E6A@CACECAEAEEE?*

2. Quality Control of raw reads

- Modern sequencing technology can generate **huge number of sequences**, but no method is **perfect**.
- We need to **assess the quality of reads** and then determine if we need to **trim or filter reads**.
- **FastQC** is a good program to indicate if raw reads are '**good**' or '**bad**'.

- Install by typing **brew install fastqc** into terminal
- FastQC is used by typing **fastqc {reads_1} {reads_2} etc.**
- **e.g. fastqc 14683_1_201.1.fastq.gz**

- FastQC outputs a html file and a .zip file. The **.html** file is the user-friendly report you can read.

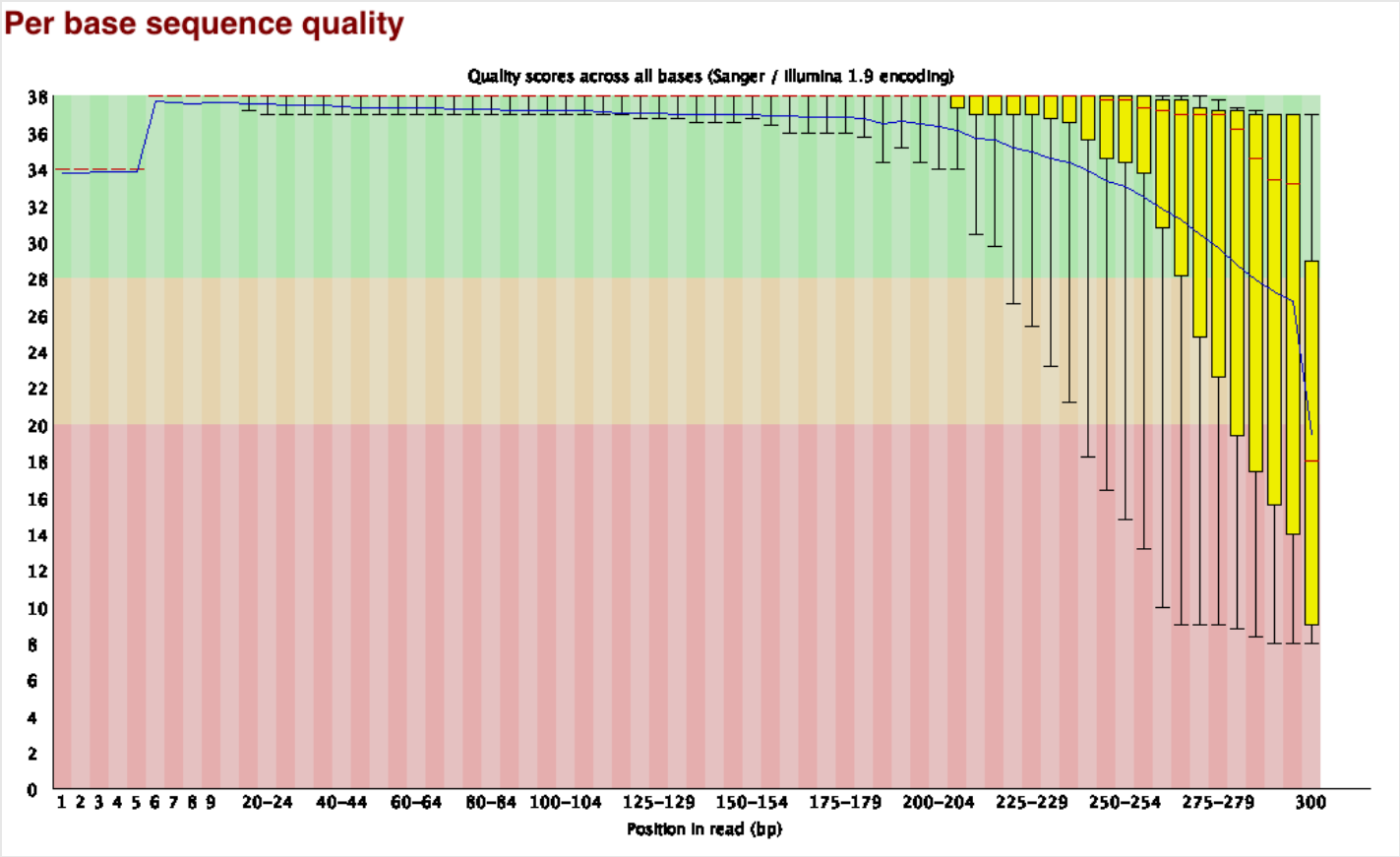
Assessing the Read Quality

- FastQC looks at **quality scores** collectively across all reads in a sample.

- X-axis shows the position of the sample.
- Y-axis shows the quality scores.

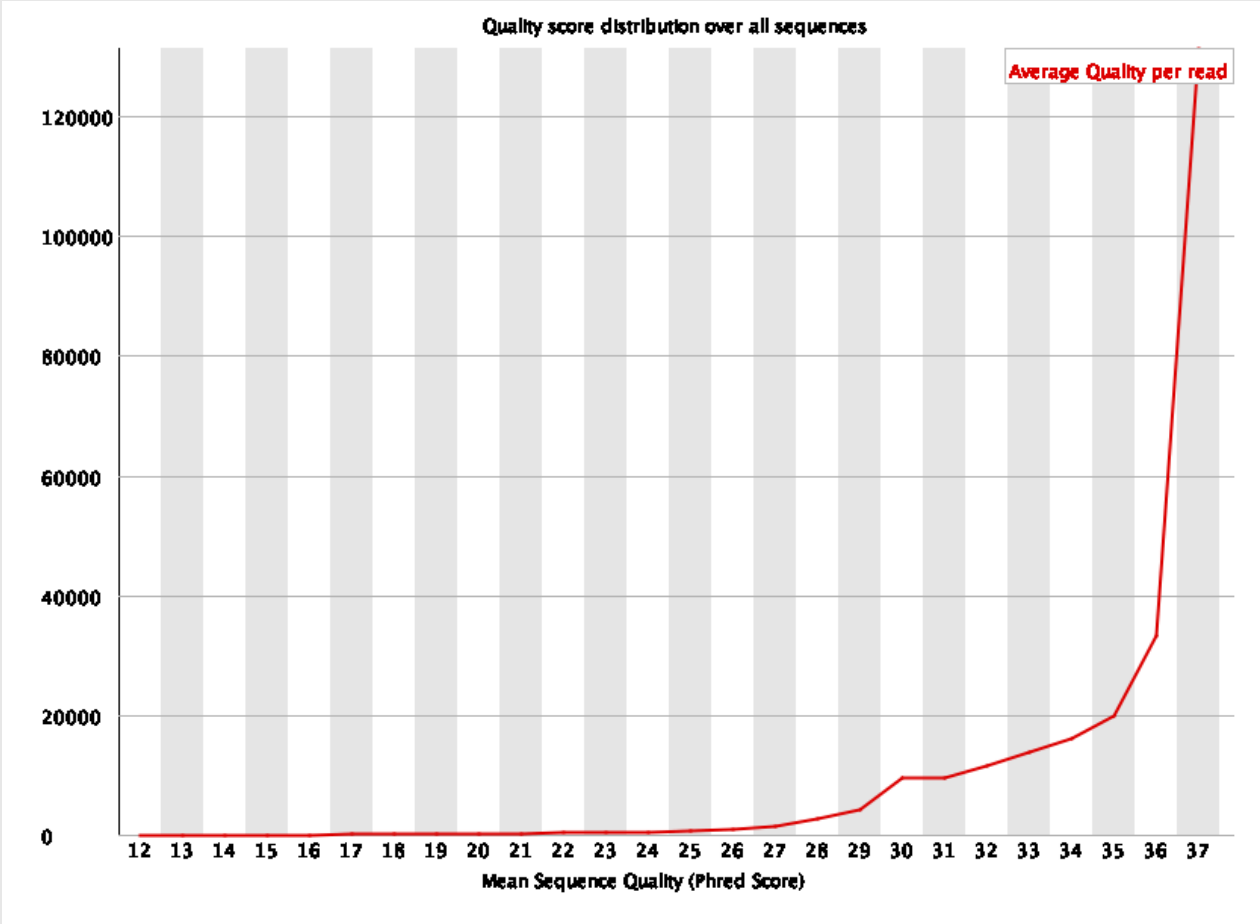
- Higher score = better base call.
- The quality of the 1st 5-7 bases is often lower due to signal decay during a run.

- Where **Phred scores** (quality scores) are <20, we consider trimming them.



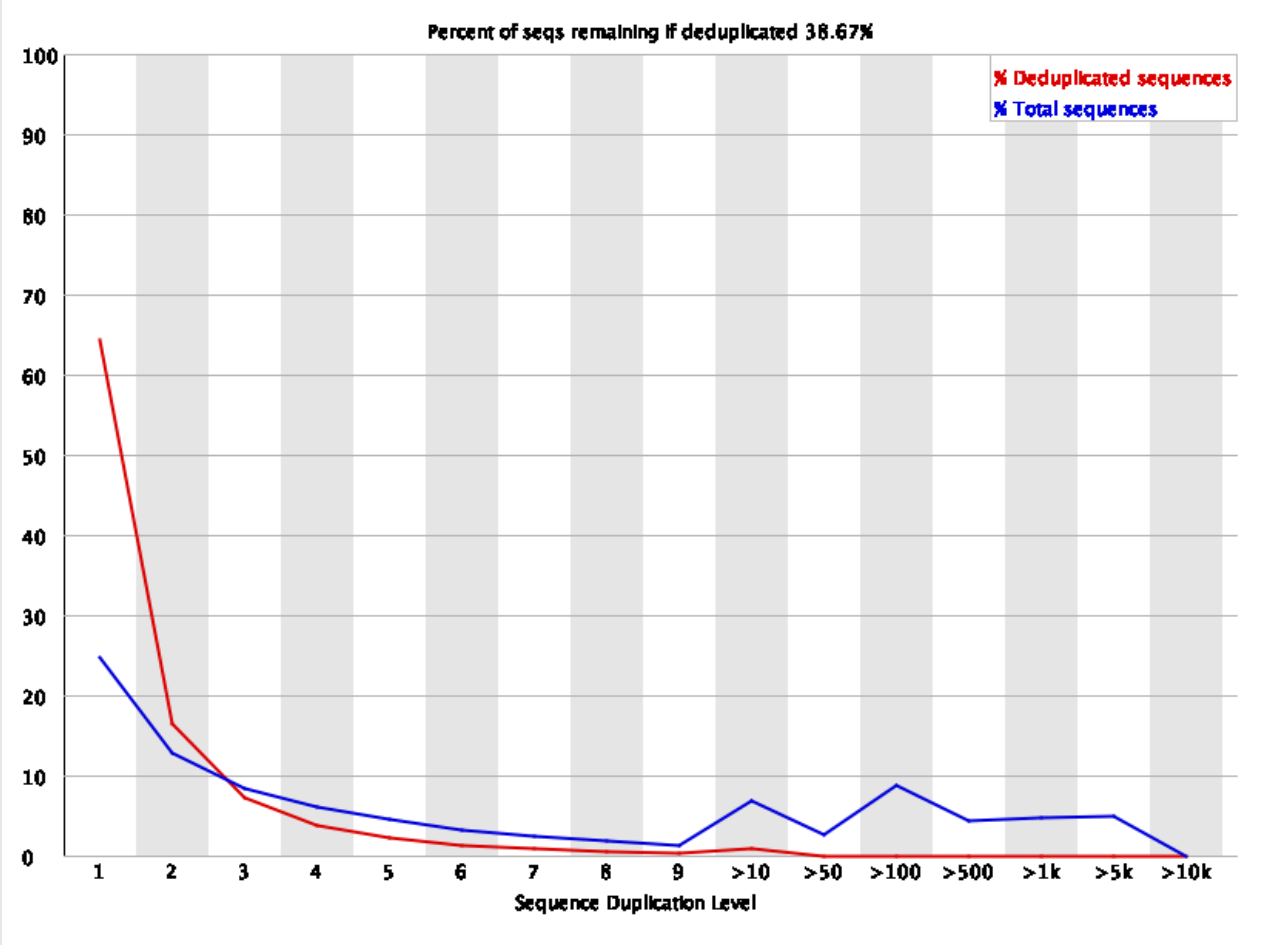
Per sequence quality scores

- A plot of the **average quality score** of all reads. This example shows a good quality score as the peak rise exponentially.
- Where he mean quality score is <27 – this means there has been an error rate of 0.2% - you will see a warning.
- If mean quality score is <20, this is a 1% error rate and is cindered by FastQC as a failure.



Sequence Duplication Levels

- In **diverse sequences**, many reads only occur **once**.
- A high level of **duplication** may be due to PCR enrichment – PCR duplicates misrepresent the **actual proportion** of sequences.
- Here, the blue line is what we actually have
- The red line is what the distribution would look like when deduplicated



Over-represented sequences

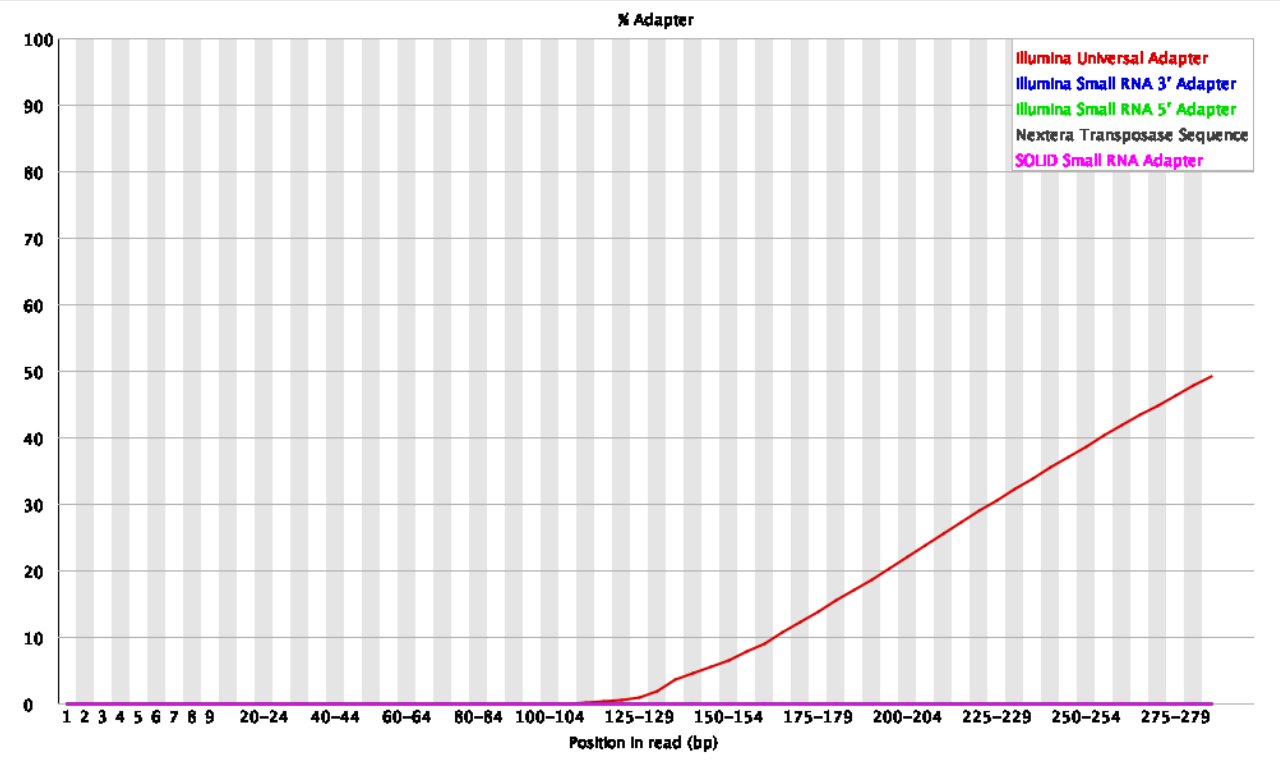
- No one sequence should make up a large proportion of the sequences.
- FastQC shows all sequences which make up **>0.1%** of the total.
- FastQC also looks at these sequences and matches it to a **known contamination database**.
- These sequences may be:
  - highly **biologically significant**
  - indicate a **contaminated library**
  - suggest that sample is **not very diverse**

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
AGCCCGGGAGCTCTCTGGCTGCTTAGGGAACCCACTGCTTAAGCCTCAAT	7271	2.779774284316124	No Hit
CCTCCAATTCCTCCTATCATTTTTCGCCTTTTAATACTGACGCTCTCGCA	5827	2.227718987032053	No Hit
AGCCTGGGAGCTCTCTGGCTGCTTAGGGAACCCACTGCTTAAGCCTCAAT	3169	1.211539637876193	No Hit
CCTCCAATTCCCCCTATCATTTTTCGCCTTTTAATACTGACGCTCTCGCA	1820	0.6958037680450208	No Hit
CCTATGGCAGGAAGAAGCTCCGGATGGAGGCCAGCTAAGCCACTCGGGGC	1647	0.6296641791044776	No Hit
TTAAAAGAAAAGGGGGGATTGGGAGGAGGGCAGAGGCACAGCGTTGAGAA	1326	0.5069427452899438	No Hit
TTAAAAGAAAAGGGGGGATTGGGGGGTACAGTGCAGGAGAAAGAATAATA	1297	0.49585576217274285	No Hit
AGCCCGGGAGCTCTCTGTGGCCTCTTTTATTAAGGCGTGAATCGCATCCC	1169	0.44692011255199415	No Hit
TTAAAAGAAAAGGGGGGATTGGGCAGCAGAGAACCTGTCCCAGTGAGGCC	1146	0.4381269880107659	No Hit
TTAAAAGAAAAGGGGGGATTGGGGGGTACAGTGCAGGAGAGAGAATAATA	1079	0.4125122339124052	No Hit
AGCCCGGGAGCTCTCTGGGACTTCACTGGTAATGTCACTAAGCCAAGTCA	993	0.3796335943234646	No Hit
CCTCCAATTCCTCCTATCATTTTTGGTTTCCATTTCCCTGGCAAATTTAT	894	0.3417849278199168	No Hit
CCTATGGCAGGAAGAAGCGGGCCCTCCTGGTGAACGCAGGCCTGGCCTTA	852	0.3257279177881086	No Hit
TGGCTTGTACCGTCAGCGCTGCGGAGATCTACAATGAGGTCCTCAGGTGG	825	0.31540555419623195	No Hit
CTTTTATGCAGCTTCTGAGGGCTCTACCAACGAACTGAAAGGAGGATGAG	806	0.30814166870565207	No Hit
CCTCCAATTCCTCCTATCATTTTATGTACACAATAGAGAGTTGCTACTGT	736	0.2813799853193051	No Hit
AGCCCGGGAGCTCTCTGACCTCTAAATTCTTATAAGTCAGAGAGTGACTG	735	0.280997675556643	No Hit
CCTCCAATTCCTCCTATCATTTTTGGTTTCCATTTCCCTGGCAAGTTTAT	650	0.2485013457303646	No Hit
CTTATATGCAGCTTCTGAGGGCTCTACCAACGAACTGAAAGGAGGATGAG	648	0.2477367262050404	No Hit
TGGCCTGTACCGTCAGCGTATCCCCCAGGTTTTGGGTTATCTGCTATAGG	620	0.23703205285050158	No Hit
CCTCCAATTCCTCCTATCATTTTTCACCACACAACTGCCAGTGTGGTCTT	598	0.2286212380719354	No Hit
TGGCCTGTACCGTCAGCGTCAATGACGCCCGCGCCCATAGTGCTCCCGGCT	591	0.2259450697333007	No Hit
TTAAAAGAAAAGGGGGGATTGGGGACTGAGAGAGGGTATGGACTCAGGTC	545	0.20835882065084413	No Hit
TGGCTTGTACCGTCAGCGTTAATGACGCCCGCGCCCATAGTGCTCCCGGCT	543	0.20759420112551993	No Hit
TTAAAAGAAAAGGGGGGATTGGGCTGGCCCTGGTAACAAAAGCCATCCGG	539	0.20606496207487154	No Hit
TTAAAAGAAAAGGGGGGATTGGGGGGAAGACTGGAGGTGACCATCTTGGG	516	0.19727183753364325	No Hit
CTTTTATGCAGCATCTGAGGGCTCTACCAACGAACTGAAAGGAGGATGAG	514	0.19650721800831905	No Hit
AGCCTGGGAGCTCTCTGGGACTTCACTGGTAATGTCACTAAGCCAAGTCA	497	0.19000795204306337	No Hit
TTAAAAGAAAAGGGGGGATTGGGGGTTGGGAGATGCAGAAGGAGAGTGAG	477	0.1823617567898214	No Hit
TTAAAAGAAAAGGGGGGATTGGGAGGAAAGTCAGTGCATTTCAGACAAGC	476	0.1819794470271593	No Hit

Adapters

- Adapters are **oligonucleotides** which can be attached to the ends of **DNA or RNA** recognised during sequencing.
- The plot shows the cumulative percentage of reads with **adapter sequences** at each position.
- Only **adapters specific** to the library type are searched.
- Ideally Illumina sequence data should not have any adapter sequence present.
- These **should be removed** as they can interfere with downstream analyses.



Filtering and Trimming Reads

- Any issues such as high duplication, low-quality sequences, inclusion of adapter sequences etc. can introduce bias in downstream analyses. Sequences must be treated to reduce that bias. Both steps can be completed using Trimmomatic.
- Cutting/**Trimming**/masking sequences
  - from low quality score regions
  - beginning/end of sequence
  - removing adapters
- **Filtering** of sequences
  - with low mean quality score
  - too short
  - with too many ambiguous (N) bases

Trimmomatic

- A program which can remove Illumina adapters and remove low-quality sequences all in one go.
- To install:
  - **cd Downloads/**
  - **wget http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/Trimmomatic-0.39.zip**
  - **unzip Trimmomatic-0.39.zip**
  - **cd Trimomatic-0.39.zip**
  - **mv /adapters/* .**
- To run:
  - **java -jar trimmomatic-0.39.jar PE -threads {number_threads} -phred33 {forward_read.fq.gz} {reverse_read.fq.gz} –baseout {examplename} ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36**
- This outputs 4 files: two unpaired reads and two paired reads – the ones you want are called {example}1P and {example}2P.

3. Mapping reads to a reference genome

- Once we have good quality reads, we can assemble these into a genome:
  - de novo
  - map to reference
- Map to reference using **BWA** (Burrows-Wheeler Aligner)

- Open terminal
- Type **sudo apt install bwa**
- Download **reference genome** fasta you wish to assemble to e.g. HXB2 or subtype-specific reference
- Type **bwa index {reference genome}**
- Type **bwa mem {reference genome} {forward read} {reverse read} –o eg.sam**

```
@SQ     SN:K03455|HIVHXB2CG     LN:9719
@PG     ID:bwa  PN:bwa  VN:0.7.17-r1188 CL:bwa mem K03455.fasta 14683_1_20_1.fastq.gz 14683_1_20_2.fastq.gz -o 14483.sam
MS8_14683:1:1101:1856:12754#20  77      *       0       0       *       *       0       0       CCCTTTCTAAGGTATTATTCCAGACCCTTCTAGTAAT
GTTACAATGTGCTCGTCTTATGTCTCCTATTATGTTCTGTGCATAGAATACCTGTCCTGGTCCTATCCTTACACTTTTTCTTGTATTATTGTTGGGTCTTGCACACGTGATTTCTACAGATTCATTAAGATGT
ACTATTATTGTCTTAGCATTGTTTGATATATTTTCAGATCTAATTATTATATCTTTTTCTGCTAGGCTACCATTTAACAGTAGTTGAGTTGACACCACTGGCTTAATCCCATGTGTACATTGTACTATGC
 @@@CCCEF9-,,C,CE9ECCE9<,C@@CC<E,,C,,C,CC<C<,C,C,6CC7@@ED9C,C@CC@F,CC,C<CECE,C,<,E9,,,<,CCC,CECF<,:CCF9CCEEFCE9CEEFEFDEEEE9EE9FF9EE8,
+B=EE,,,:?A7A8,?ED?E9A<<,CAA9EF9,;,C<A;AF;AE,EE9EAFF99,@=@EEEFF9>ADADEGGC;D,@=F9=DDADD:FCDEGFFGGGGF8EA,66A??DDD?FCD==+C?9=F7;*BF77;=:
5??FF*69AFCFCEA@A6E6A@CACECAEAEEE?*     AS:i:0  XS:i:0
MS8_14683:1:1101:1856:12754#20  141     *       0       0       *       *       0       0       ATTCAATGGGACAGGACCATGCAGTAACGTCAGCATA
GTACAATGTACACATGGGATTAAGCCAGTGGTGTCAACTCAACTACTGTTAAATGGTAGCCTAGCAGAAAAAGATATAATAATTAGATCTGAAAATATATCAAACAATGCTAAGACAATAATAGTACATCTTA
ATGAATCTGTAGAAATCACGTGTGCAAGACCCAACAATAATACAAGAAAAAGTGTAAGGATAGGACCAGGACAGGTATTCTATGCACAGAACATAATAGGAGACATAAGACGAGCACATTGTAACATTAC
 -8A@--<--,,;,,,,:9,;,6,,=,,,,;,,,,C,,<,<,,<,C,;,6,C,,,,;C,,,6;,,;,,<,;@,,;CC,,:C,99,<6,,,C,,<,,55C,,95,,,,,,,:,A,,C,,9A,,,@EFA9,,,
A,C,AE,,,@,,=>@D9,4,4,4@9,CD9E;C9DBAEFCE8=8=ADDDFC88,@;ED+=1=8==,++===++=++@+6=+?+6+***0*59<C779*+;****895))08)0)37;77;*@76*6*0)31:<4
*:**.*0)-):6964>22(59B3;C4>4C?9EEE)     AS:i:0  XS:i:0
MS8_14683:1:1101:1980:10573#20  77      *       0       0       *       *       0       0       TTTGGAAGGACACCAGTCAGATTGGACTAGGGTCCAC
CCTGATGACCTCATTTGACCCAAATTACCTCCAAAGAAGGTCATACCCTGAAGTCCTGGGGGCGAGGACTTCAACACACCCTGCTCTGGGGAATGGGGTTGAGCCACTGCCTGAGGCACAGGACGGGGCTCCC
TGGATCACGGGGGTCAAAGCTGAAGCCATCACATCTGTCCCAGGGAGACCCTGAGCTGGCTTTGGGTAATGGGACACGGGGCCGTACGGCCCATGTGCGGGGACTGAGCCAGGGTTCGGCAGGGGTGTTG
 @CCCCCEGGGFEE@FGGGFGGGGCFDEGEFFGGGDECF@FGCFFEFCGGFFGGFGGFC@FEFGGDEFFGFCCEFGGGGFEGGCCEE<EEGGGDGGGFG>FGEEGEGGGGEFBEEFEEGGGCFEGGGGE7BFF
GGGGGGCFCCCFGGGFGGCFFE@FCGGFGGDGGGGGECEGGGG:FBFGCFEGEFFCCFEGGFGGCFGGGGGCCGGGFF9BFFGFGE==FFF6FGGGGDGCFGGDGGGGFGGDDFFGGCC>DGD5DFD375<FF
F=?EDEFFFF:4@(54>F:<((302-70(((-3:(     AS:i:0  XS:i:0
```

- The output file we have requested is a **SAM file** (Sequence Alignment Map)
- Tab-delimited file of alignments of short reads mapped against a reference sequence.
- Read Name → SAM flag → contig name → mapped position of base 1 of a read on the reference sequence → mapping quality → CIGAR string → name of mate → position of mate → template length → read sequence → read quality → Other information in TAG:TYPE:VALUE format.

Converting SAM to BAM file

- SAM files are extremely large (but readable by humans). We store alignments as BAM files (Binary Alignment Map) which is compressed version of a SAM.

- Install samtools:
- **sudo apt install samtools**
- To change a SAM to a BAM type: **samtools view -h -b -S {example.sam} > {example.bam}**

- When aligning to a reference genome, many reads may not be mapped to the reference.
- We can remove these to further reduce to BAM size.
- Type **samtools view -b -F 4 {example.bam} > {mapped.bam}**

4. Sorting and Indexing a BAM File

- BAM files can be sorted by read name or by coordinates (chromosomal position of an alignment).
- Some downstream analyses require that BAM files be **sorted** and **indexed**.
- We can do both of these things using **samtools**.
- Samtools sort – sorts alignments by leftmost coordinates
- Samtools index – indexes a co-ordinate sorted BAM for fast random access

- To sort, type: **samtools sort {input.bam} –o {sorted_output.bam}**
- To index, type: **samtools index {sorted_output.bam}**

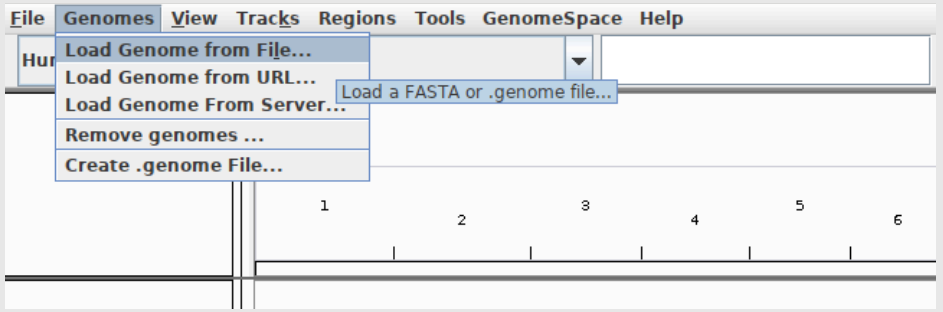
5. Viewing your sorted BAM file

- After creating the BAM and then sorting and indexing it, we can use a viewer to identify any areas of poor coverage.
- A well-known BAM viewer is the Integrative Genomics Browser (IGV).
- To install:
  - Download the Linux archive: go to Downloads/ in terminal and type
  - **wget https://data.broadinstitute.org/igv/projects/downloads/2.7/IGV_Linux_2.7.2.zip**
  - **unzip IGV_Linux_2.7.2.zip**
  - **cd IGV_2.7.2**
  - **sh igv.sh**

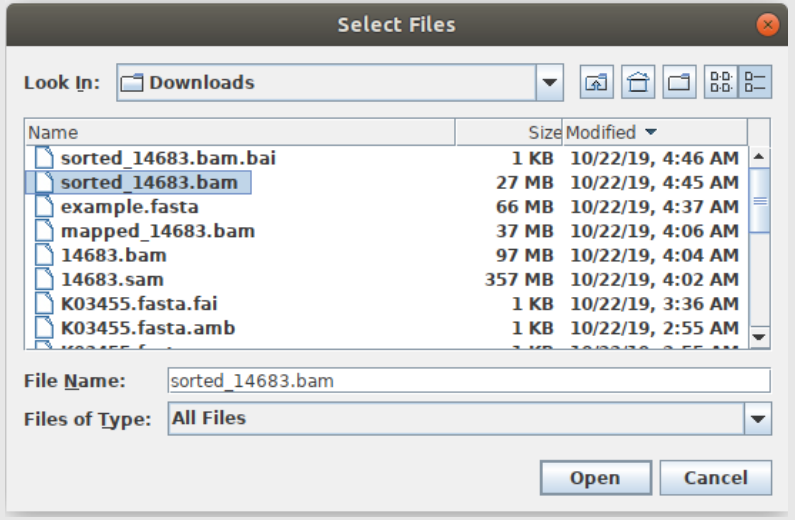
- This will open the IGV viewer.

How to use IGV to view alignment

- You must first load the reference genome that the BAM file was made from. Click Genomes → Load genome from File… and select the fasta file that we used earlier.



- You can then load the sorted BAM file, and index file. Click File → Load from File… and select the sorted BAM file.




File Genomes View Tracks Regions Tools GenomeSpace Help
K03455.fasta
K03455|HIVHXB2CG
9,656 bp
bp
1,000 bp
2,000 bp
3,000 bp
4,000 bp
5,000 bp
6,000 bp
7,000 bp
8,000 bp
9,000 bp
sorted_14683.bam Coverage
[0 - 48867]
sorted_14683.bam

6. Post-alignment Quality Control

- Once you have eyeballed the alignment and are happy with it, you can perform some additional QC to see if there are any issues.
- We can do this by using FastQC again.
- Type:
  - **fastqc {input.bam}**
- After this has completed, you can open the .html file again to see the report.

7. Downstream Analysis

GENERATE PHYLOGENETIC TREES

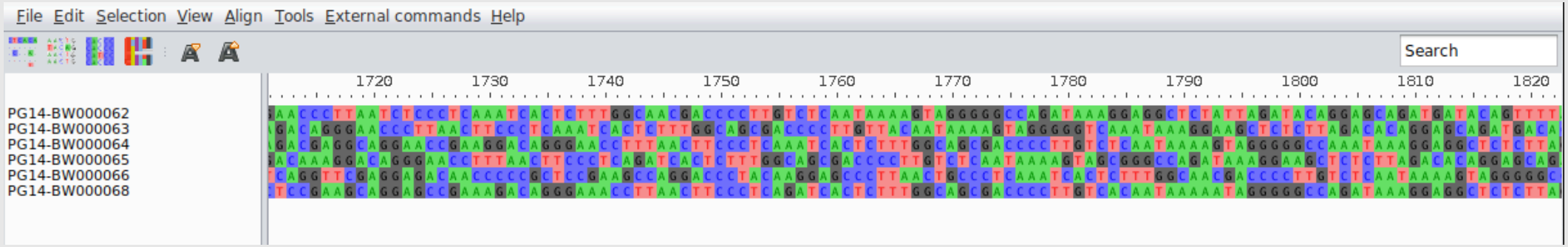
ANALYSE SEQUENCES FOR DRUG-RESISTANCE MUTATIONS

ANCESTRAL RECONSTRUCTION

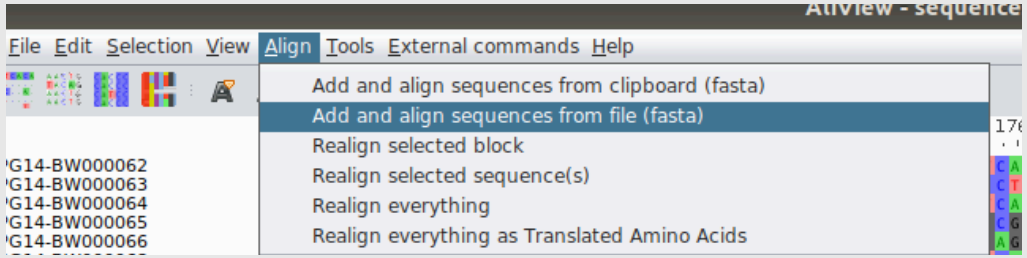
Phylogenetic Trees

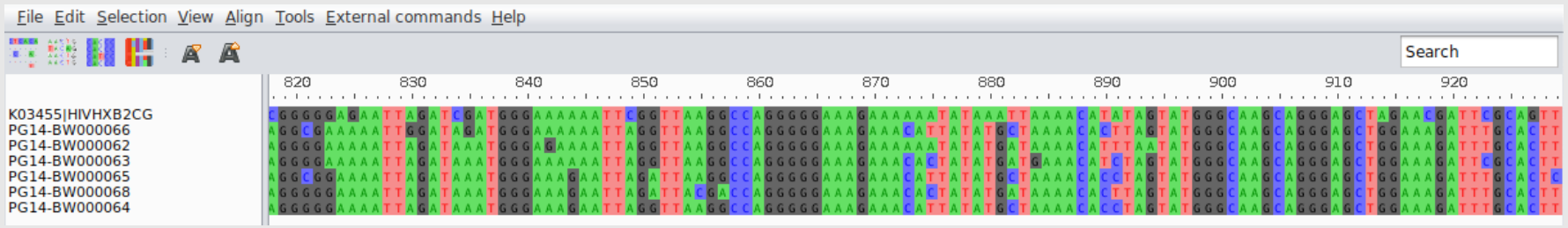
- If using Phyloscanner – input is the BAM files you have made.
- If creating a maximum-likelihood or neighbourhood-joining tree, these use the **consensus sequence** of the BAM files. PANGEA create a consensus sequence using a tool called **SHIVER**.
- Researchers who are **accredited** by PANGEA can request these sequences which saves you from having to make them yourselves.
- You can then use these to create a multiple sequence alignment (MSA):
  - Place all files into a folder
  - concatenate all sequences into a single file; type: **cat * > sequences.fasta**
- You can then install a sequence viewer – I recommend Aliview. To install:
  - **wget https://ormbunkar.se/aliview/downloads/linux/linux-version-1.26/aliview.install.run**
  - **chmod +x aliview.install.run**
  - **sudo ./aliview.install.run**
  - **aliview**

- After installing and typing ‘aliview’ into the terminal, AliView will open!
- Open your concatenated consensus sequences by clicking File → Open File and then selecting your sequences.
- It will look like this:



- We need to add a reference to align everything to now
- Click on Align → Add and align sequences from file (fasta) and choose your reference sequence.
- The programme will now insert the reference genome and perform a multiple sequence alignment.



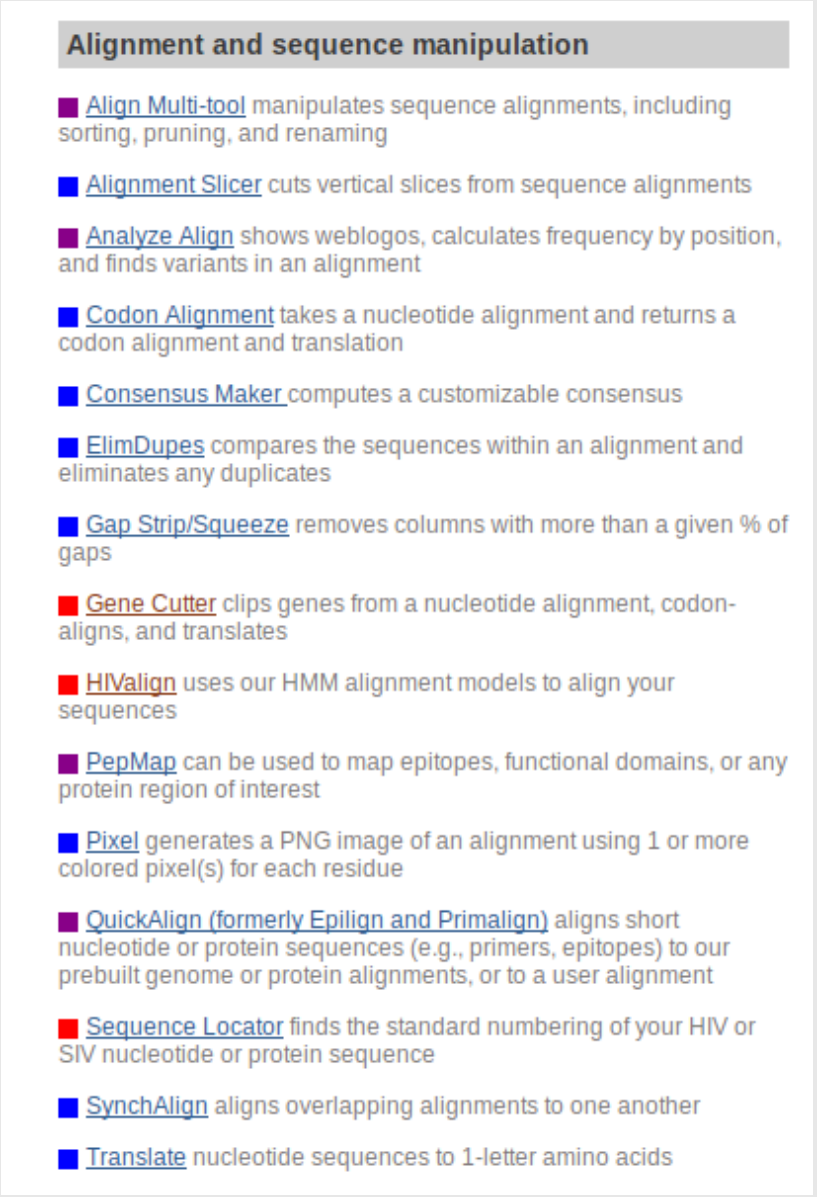


- Alternatively, you can use online tools such as HIValign or Gene Cutter from LANL: https://www.hiv.lanl.gov/content/sequence/HIV/HIVTools.html

- HIVAlign takes your unaligned fasta file and aligns to HXB2 using MAFFT.

- Gene Cutter aligns nucleotide sequences and codon aligns all coding regions. This can output a nucleotide alignment and protein alignments.

- Whichever tool you use, you'll get an aligned fasta file at the end. Use this to build trees.



Alignment and sequence manipulation

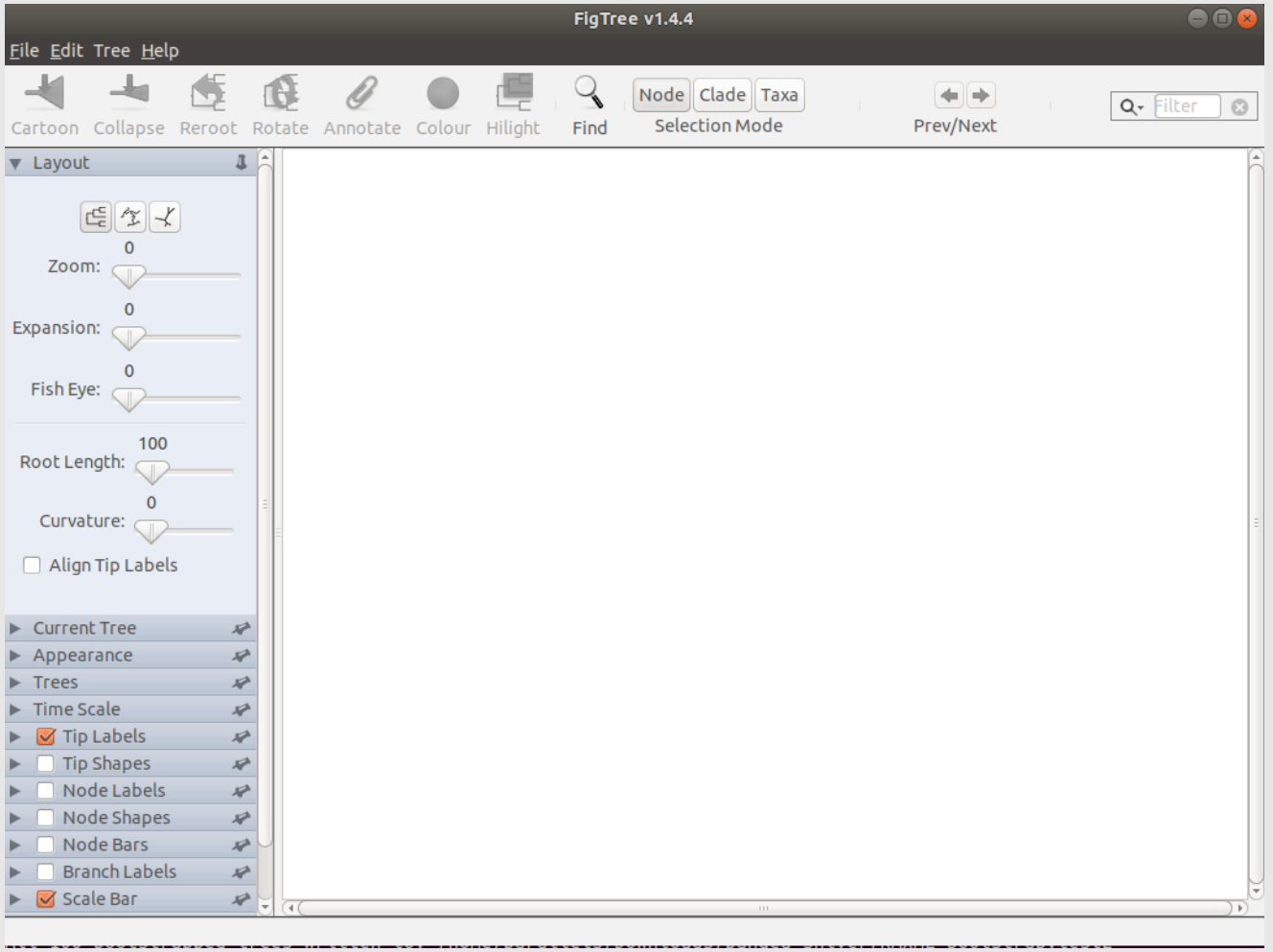
- Align Multi-tool manipulates sequence alignments, including sorting, pruning, and renaming
- Alignment Slicer cuts vertical slices from sequence alignments
- Analyze Align shows weblogos, calculates frequency by position, and finds variants in an alignment
- Codon Alignment takes a nucleotide alignment and returns a codon alignment and translation
- Consensus Maker computes a customizable consensus
- ElimDupes compares the sequences within an alignment and eliminates any duplicates
- Gap Strip/Squeeze removes columns with more than a given % of gaps
- Gene Cutter clips genes from a nucleotide alignment, codon-aligns, and translates
- HIValign uses our HMM alignment models to align your sequences
- PepMap can be used to map epitopes, functional domains, or any protein region of interest
- Pixel generates a PNG image of an alignment using 1 or more colored pixel(s) for each residue
- QuickAlign (formerly Epilign and Primalign) aligns short nucleotide or protein sequences (e.g., primers, epitopes) to our prebuilt genome or protein alignments, or to a user alignment
- Sequence Locator finds the standard numbering of your HIV or SIV nucleotide or protein sequence
- SynchAlign aligns overlapping alignments to one another
- Translate nucleotide sequences to 1-letter amino acids

Tree-Building

- Three major types of tree: **Neighbourhood-Joining** (NJ), **Maximum-Likelihood** (ML) & **Bayesian** (BEAST).
- Argument for both, NJ is **faster** but considered to be less accurate. ML is slower and uses more **complex evolutionary models**. BEAST uses markov-chain monte-carlo (MCMC) to generate rooted, **time-measured phylogenies**.
- Recommend using software called **raxml** – for building ML trees with bootstraps.
- Install by
  - **sudo apt install raxml**
- Build trees and perform rapid bootstrapping on one go:
  - **raxmlHPC –f a -m GTRGAMMA -p ${RANDOM} -x ${RANDOM} -# {number_bootstraps} -s {alignment.fasta} -n {name}**
- This will do x bootstrap searches, 20 ML searches and return the best ML tree with bootstrap values.
  - 100 bootstraps = fast tree
  - 500-1000 bootstraps publication quality.

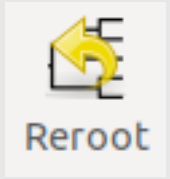
Viewing Trees

- Use a tree-viewer such as Figtree. Can also do some manipulation of trees in this programme.
- To install:
  - **sudo apt install figtree**
  - **figtree**
- Open the correct file to view your tree. Click File → Open, then select the file called RAxML.bipartitions.example_name
- An input window will open – type bootstraps and click OK.

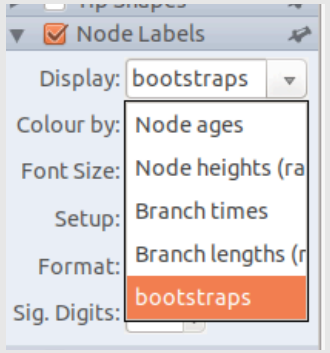


Manipulating Trees

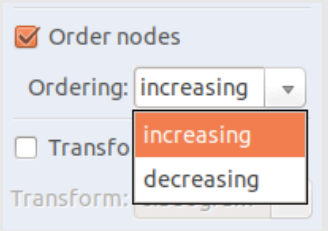
- Can do various things in Figtree:
- Re-root trees
  - Click on nodes you wish to reroot at (outgroup). Click Reroot



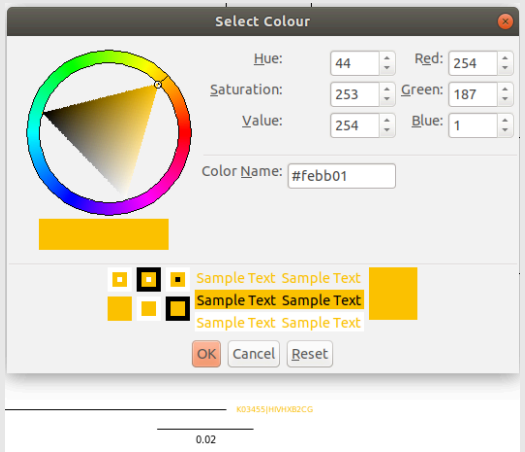
- Show bootstrap values
  - Tick Node Labels and change Display: to bootsraps



- Sort nodes by increasing or decreasing evolution.
  - Select Trees, then Order Nodes (by increasing)

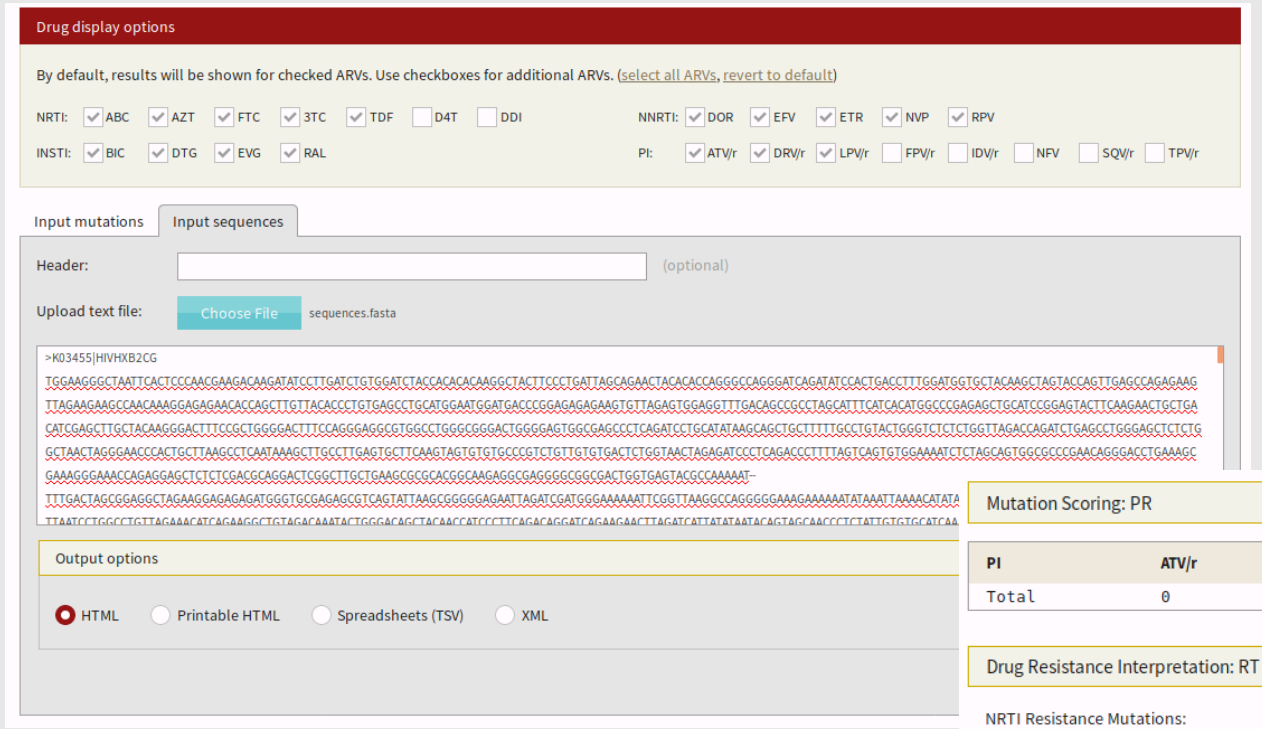


- Colour branches or Taxa or Nodes
  - Choose Node/Clade/Taxa as appropriate
  - Select the Node/Clade/Taxa you want to colour
  - Select Colour and Choose a colour!



Looking for Drug Resistance Mutations

- Easiest way is to use the **Stanford HIV database -** https://hivdb.stanford.edu/hivdb/by-sequences/
- Input the sequences you **wish to analyse**, and it gives a report which you can view on-screen or a spreadsheet that you can download, for many sequences. Limited to **1000 sequences** of up to **3000 nucleotides**.
- This tells you the resistance mutations and which drugs would be susceptible or resistant.



Mutation Scoring: PR

PI	ATV/r	DRV/r	LPV/r
Total	0	0	0

Drug Resistance Interpretation: RT

NRTI Resistance Mutations: None
NNRTI Resistance Mutations: None
Other Mutations: K122E, F214L, A272P, K277R, A376T, A400T, D460N, S468T, H483Y, K512Q, S519N

Nucleoside Reverse Transcriptase Inhibitors		Non-nucleoside Reverse Transcriptase Inhibitors	
abacavir (ABC)	Susceptible	**doravirine (DOR)**	Susceptible
zidovudine (AZT)	Susceptible	**efavirenz (EFV)**	Susceptible
emtricitabine (FTC)	Susceptible	**etravirine (ETR)**	Susceptible
lamivudine (3TC)	Susceptible	**nevirapine (NVP)**	Susceptible
tenofovir (TDF)	Susceptible	**rilpivirine (RPV)**	Susceptible

Mutation Scoring: RT

NRTI	ABC	AZT	FTC	3TC	TDF
Total	0	0	0	0	0

NNRTI	DOR	EFV	ETR	NVP	RPV
Total	0	0	0	0	0

Tools for Help

- Many of these processes can be automated using an online system called Galaxy : https://galaxyproject.org/tutorials/g101/
- This is a user-interface where you upload all of your sequences and choose the tools you want to use, and this will run on their servers.
  - Can be a little slow and steep initial learning curve
  - Means you don't have to leave computer running or worry about power interruption etc.
- A paid tool exists which is useful for many analyses including sequence building, multiple alignments, SNP calling, tree building etc: https://www.geneious.com/
  - User-friendly, no coding needed
  - Expensive - $450 per user, per year.
  - 14-day unrestricted free trial

Getting Accredited by PANGEA

To become a **PANGEA-accredited researcher** and get access to sequences, you must fill in a **PANGEA accreditation form** and attached a **certificate** of a course on **Human Subjects Research**.

If you are a member of PANGEA and are employed by MRC Uganda, Rakai, Partners/UW, Botswana/Harvard, and PopART, there is funding for you to take these courses.

Please email Lucie if unsure, some institutions have free institutional access

Once you are accredited, you can formally request data from PANGEA and do some analyses.

Any other information?

- Are there any tools or specific analyse which you want to perform?
- Send me a list of tools you may wish to use, or questions with specific analyses and PANGEA will try and help you with this ☺
- Any tools that you may wish to use can be explored in next month's webinar (14th November).