Introduction to Geneious and Databases

CPHL Nextstrain Training

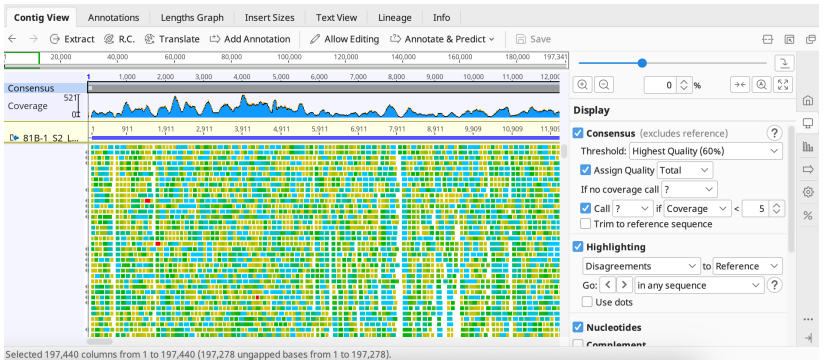
June 2025

Shaun Cross, PhD

What is Geneious?

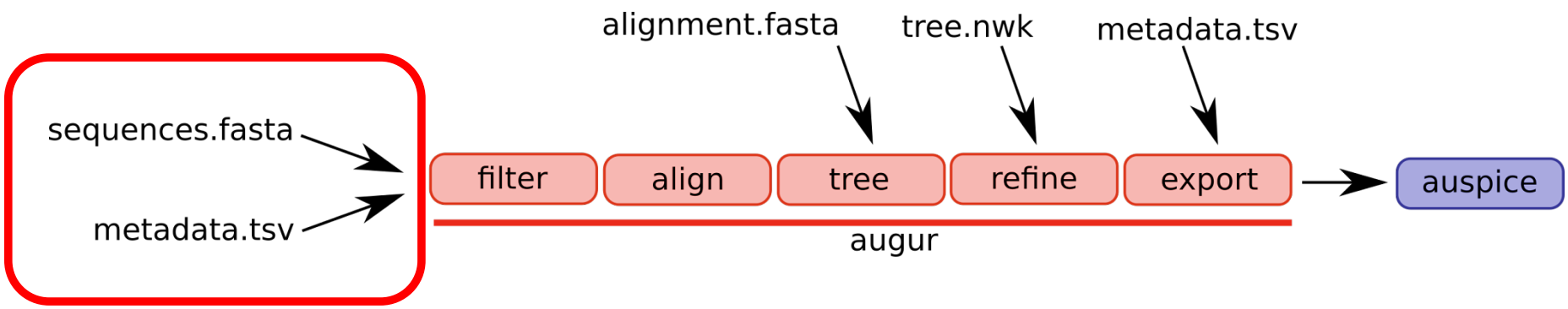
- A software platform bioinformatics software platform used for analyzing and interpreting sequence data.
- It is a GUI (Guided User Interface) that allows for raw sequence data to ‘transform’ into visualizations

Why use Geneious?

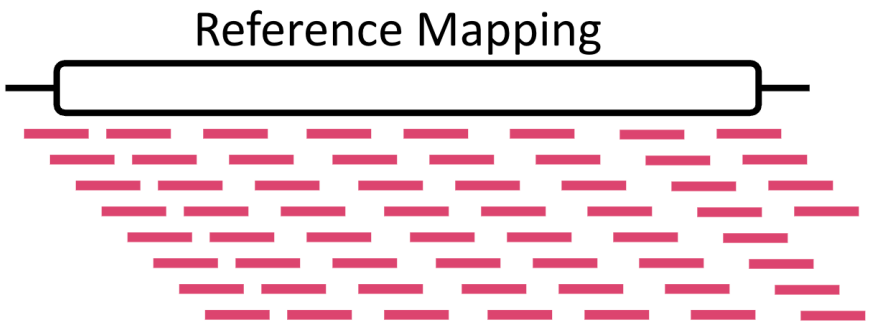


- There are A LOT of useful functions of Geneious:
  - Sequence alignments
  - SNP analyses
  - Genome coverage plots
  - Read mapping and de novo assembly
  - Phylogenetics
  - And more molecular biology tools (primer mapping/design, cloning tools, etc.)
- It makes for really great “sanity checking” steps for your sequence analyses!

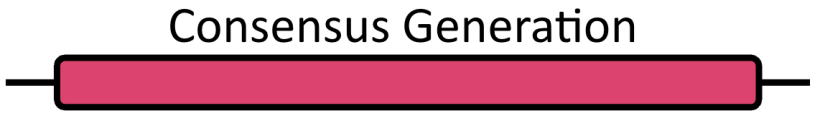
Geneious is helpful for preparing data to go into Nextstrain builds/workflows



Raw “reads” (.fastq)

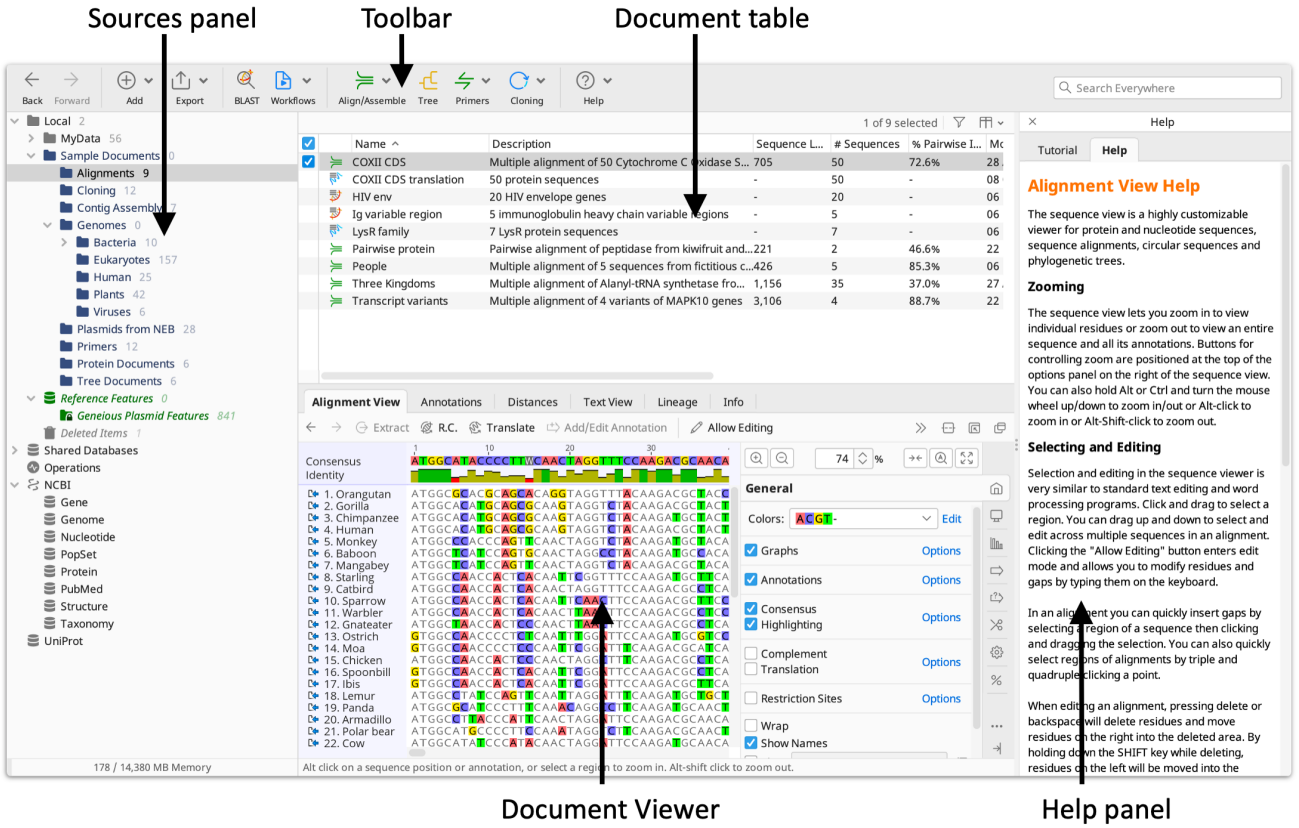


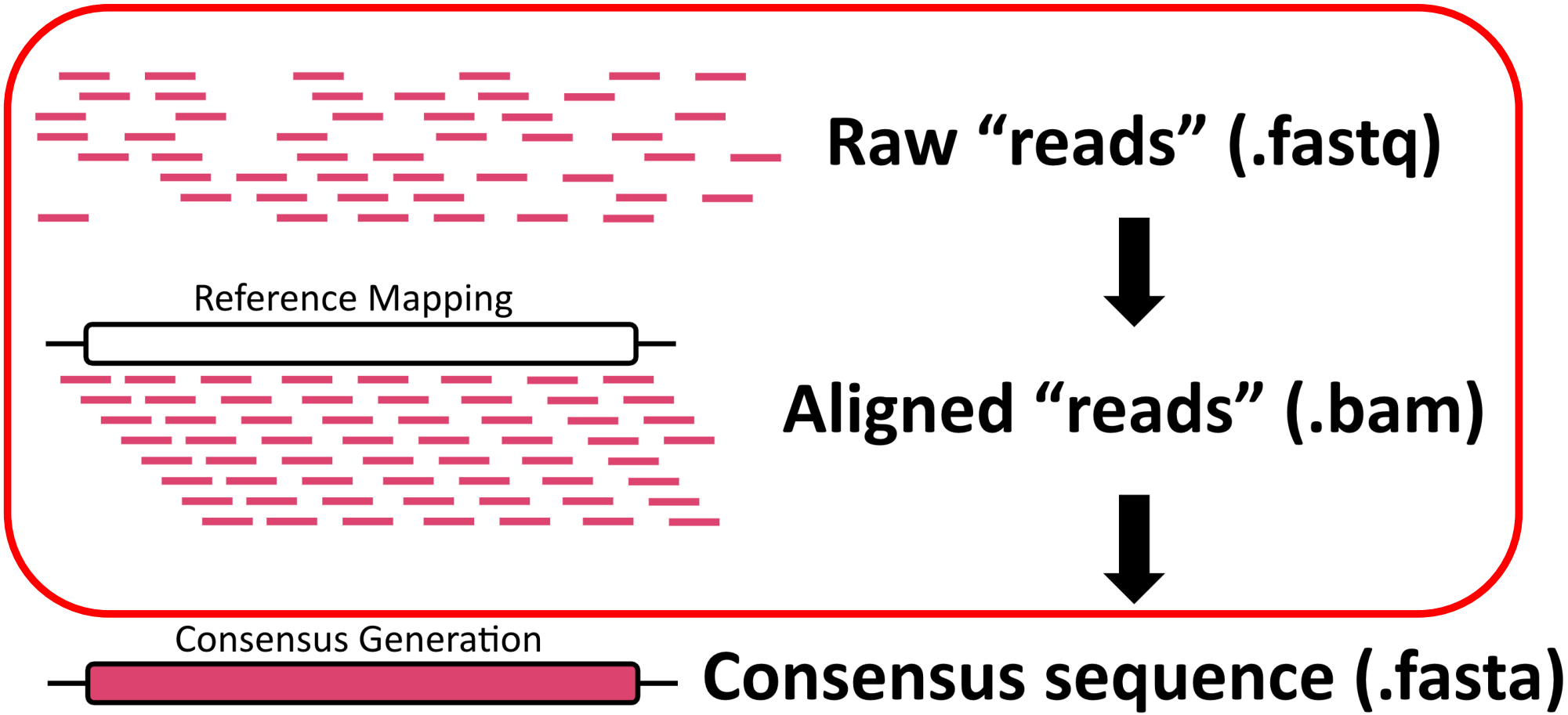
Aligned “reads” (.bam)



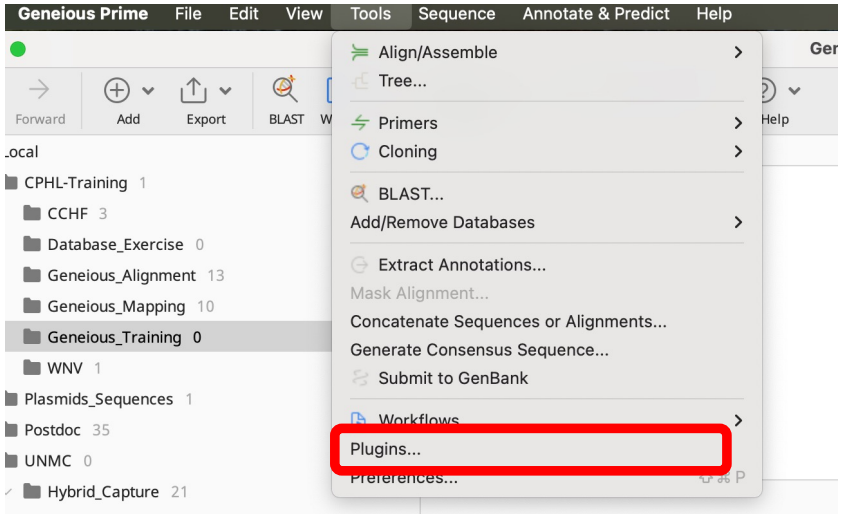
Consensus sequence (.fasta)

Let’s explore the layout of Geneious!

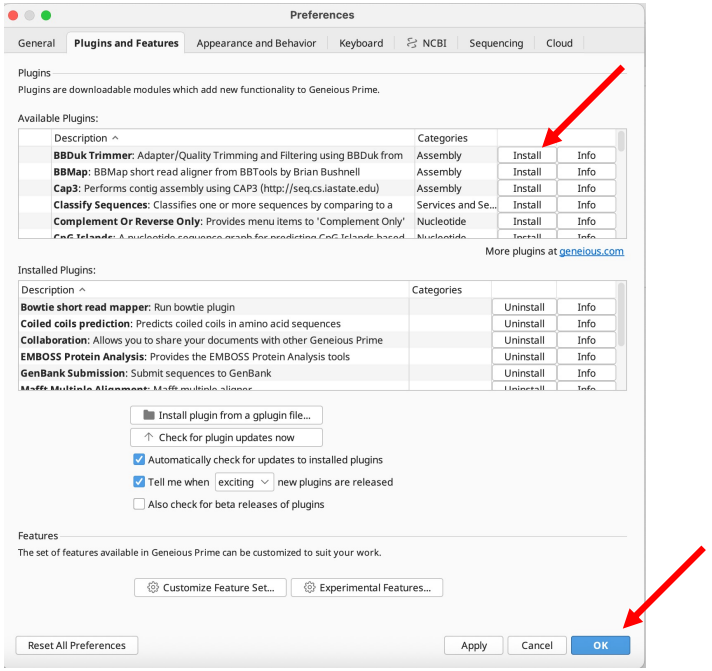



Raw “reads” (.fastq)
Reference Mapping
Aligned “reads” (.bam)
Consensus Generation
Consensus sequence (.fasta)

Visualize raw FASTq Reads and Trim and Quality Control

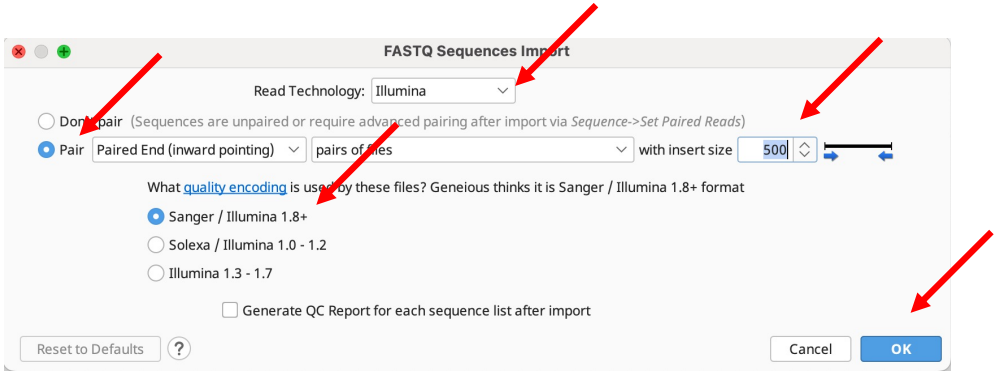


1. Select Tools, then Plugins...

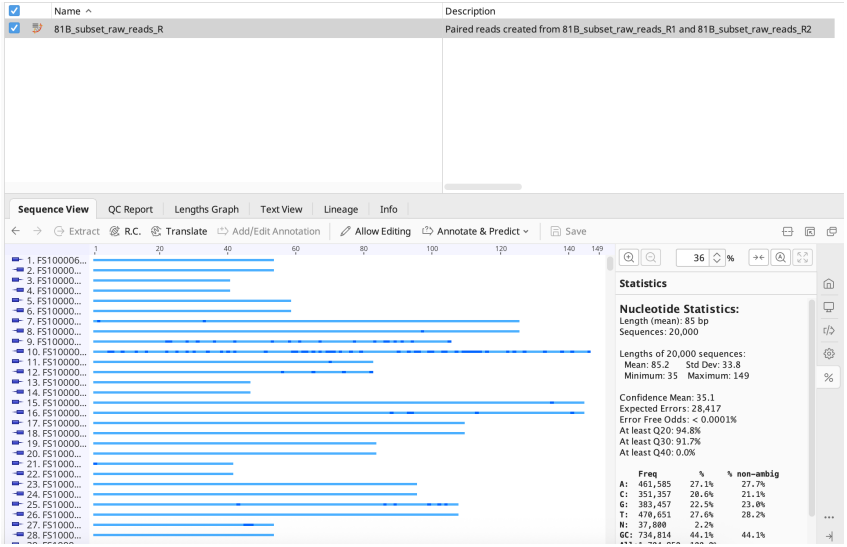


2. Install BBDuk Trimmer, then OK

Visualize raw FASTq Reads and Trim and Quality Control

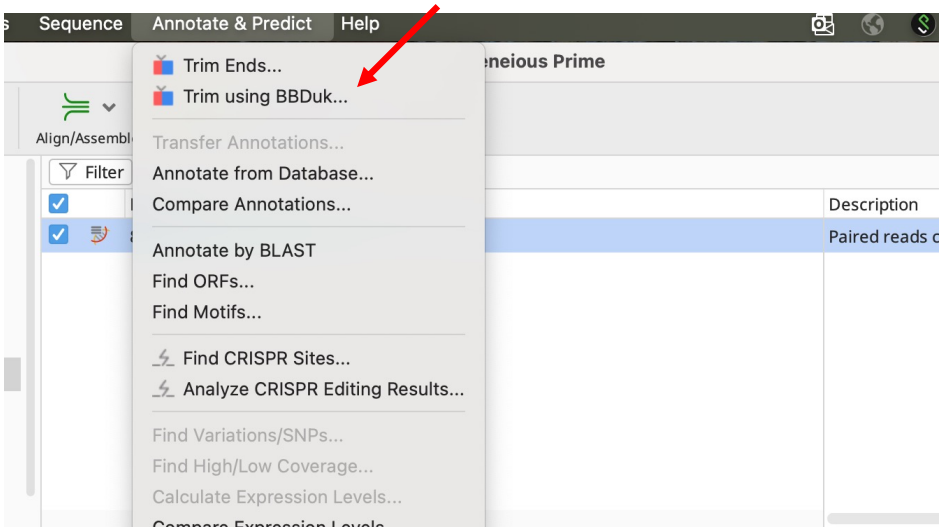


3. Drop in your raw FASTq reads (drop in R1 and R2 simultaneously if working with paired-end data). A window prompt will appear. We can keep the insert size as default 500, even if shorter—but feel free to adjust if you know the true insert size. Make sure it is using latest Illumina quality and then select OK.

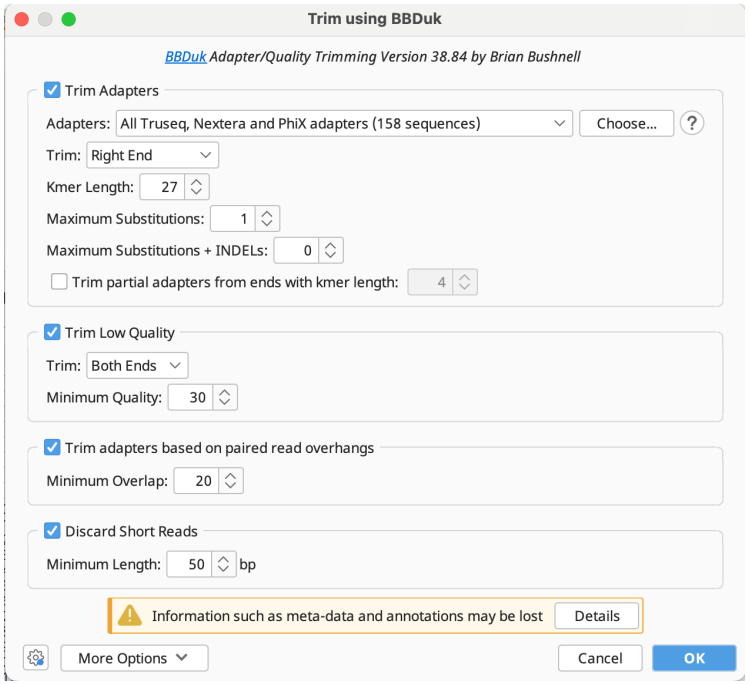


4. Visualize your raw read data with the Document Viewer section.

Visualize raw FASTq Reads and Trim and Quality Control

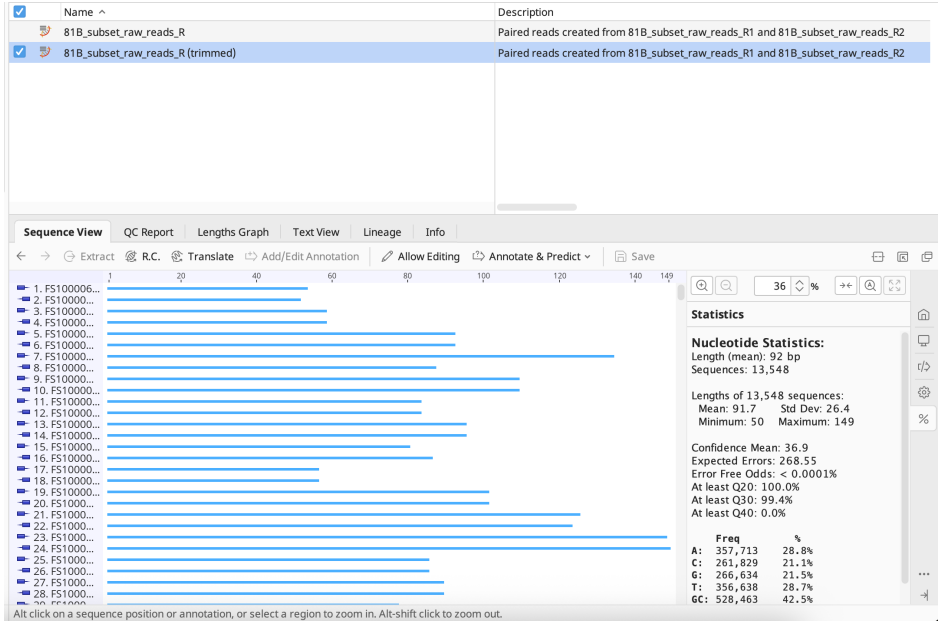


5. Make sure your file of reads is selected. Then go to Annotate and Predict and select “Trim using BBDuk”



6. In the options window, select all options. You can adjust the scoring, adapter overlap, and minimum length, but the options shown here are typically OK standard.

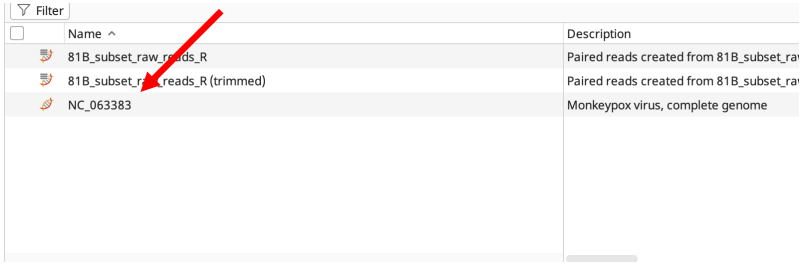
Visualize raw FASTq Reads and Trim and Quality Control

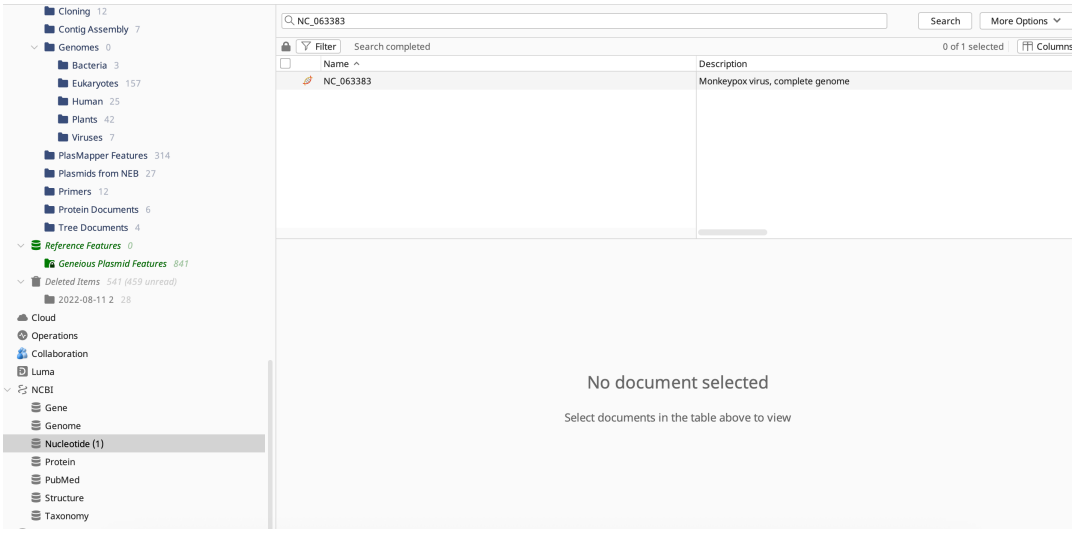


7. Now you can visualize your trimmed data!

***Note: Not usually recommended to use Geneious for trimming and quality control as computational power for large datasets exceeds Geneious limitations

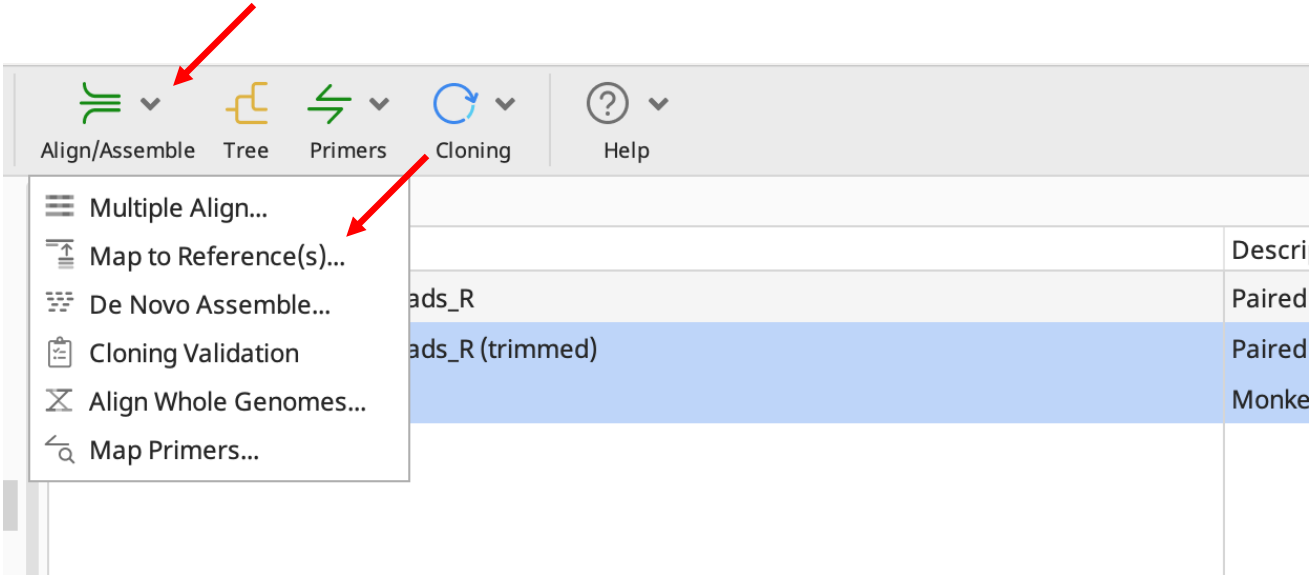
Align genomes to reference sequence





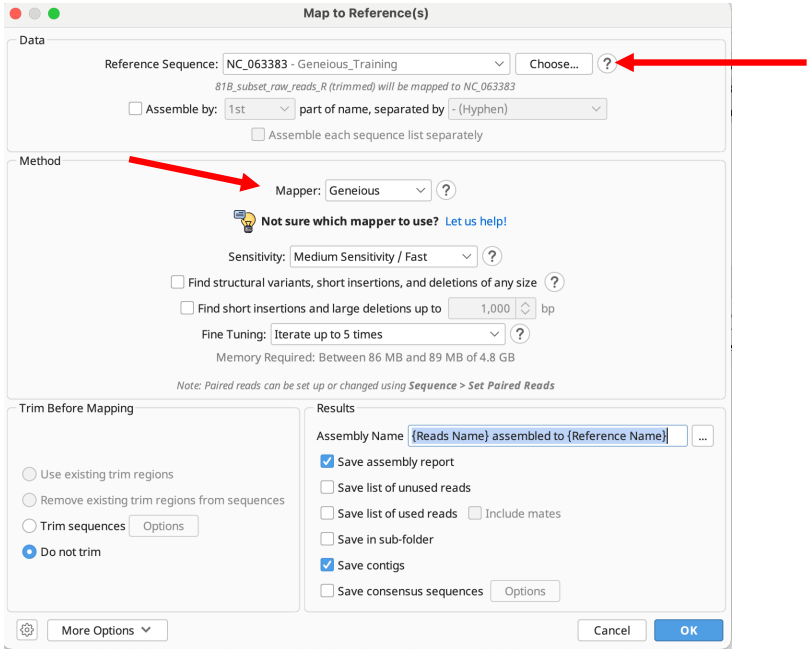
1. Drop in your reference FASTA file of interest. Preferrably, download it with through the NCBI database link and then drag and drop the NCBI file to the working folder.

Align genomes to reference sequence



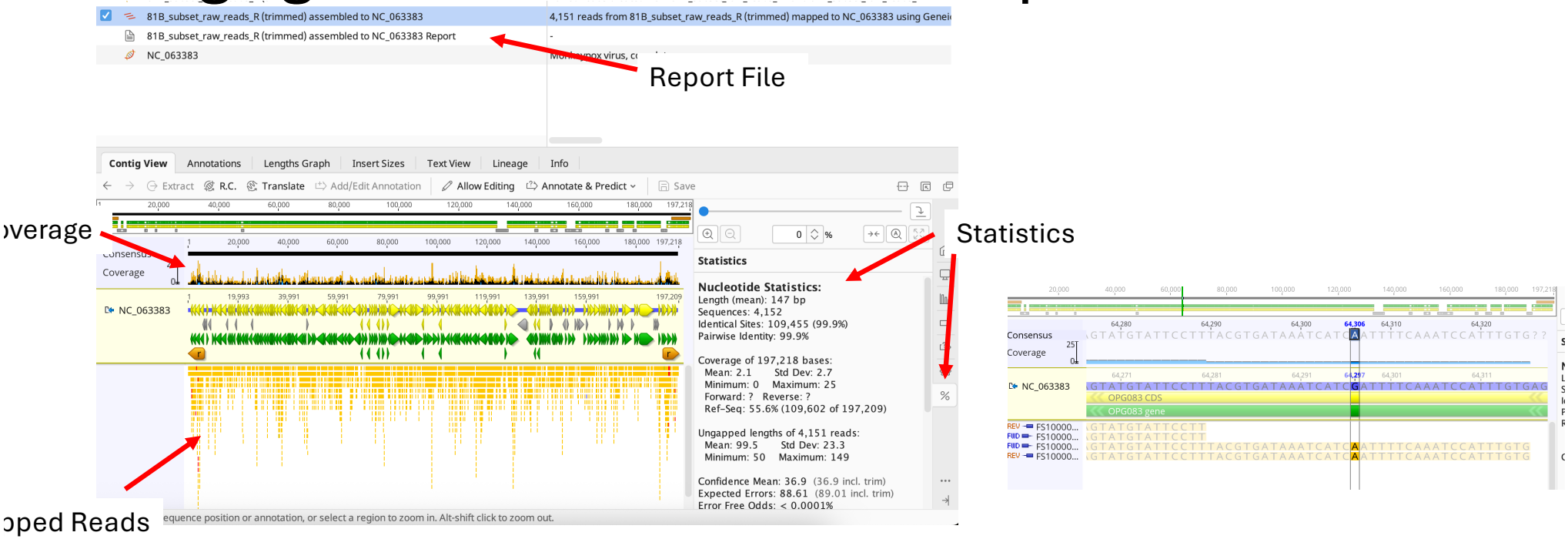
2. Select the reference genome and the reads (trimmed/QC'ed). Select Align and Assemble from the Tool bar. Then select Map to Reference(s)...

Align genomes to reference sequence



3. Make sure the right reference sequence was selected. Then choose the mapper of your choice, and adjust the options as needed. For a ‘quick and simple’ approach, I typically use the Geneious default. However, I usually run BWA or bowtie2 when working from the command line and in my bioinformatic pipelines.

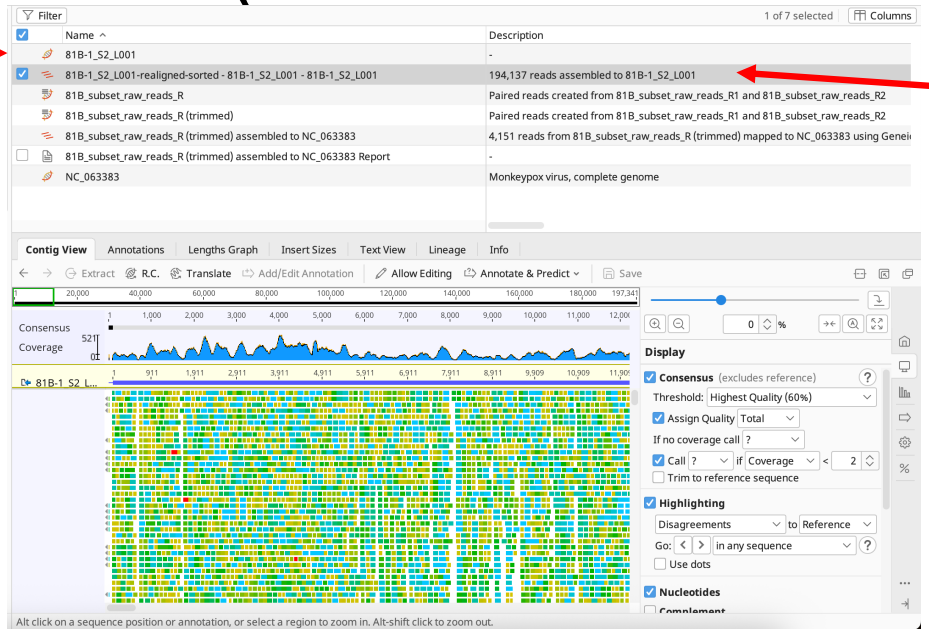
Align genomes to reference sequence



4. You can now visualize how the reads map to the genome, identify coverage, and get statistics on the data. A second “Report” file will tell you how many reads mapped as well as other details. You can zoom in as needed to visualize SNPs

You can also add in already aligned sequences! (BAM and SAM files)

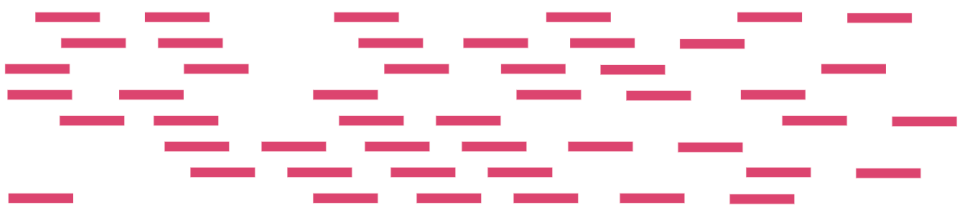
FASTA file used as reference



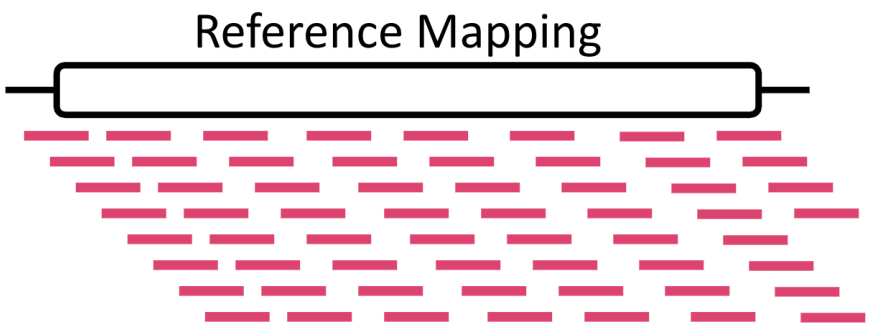
.bam file

1. Drop in the FASTA file used for alignment. Then with that file highlighted, drop in your .bam (or .sam) file. In this example, I wanted to visualize mapping reads against the generated consensus sequence. You will get the similar view as before showing coverage, mapped reads, and statistics.

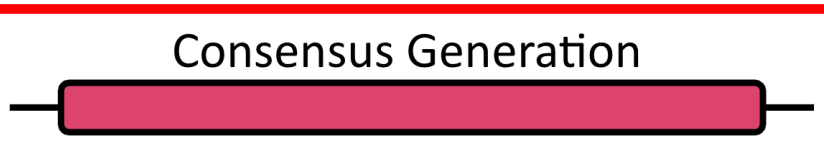
Generate a consensus sequence from aligned reads.



Raw “reads” (.fastq)

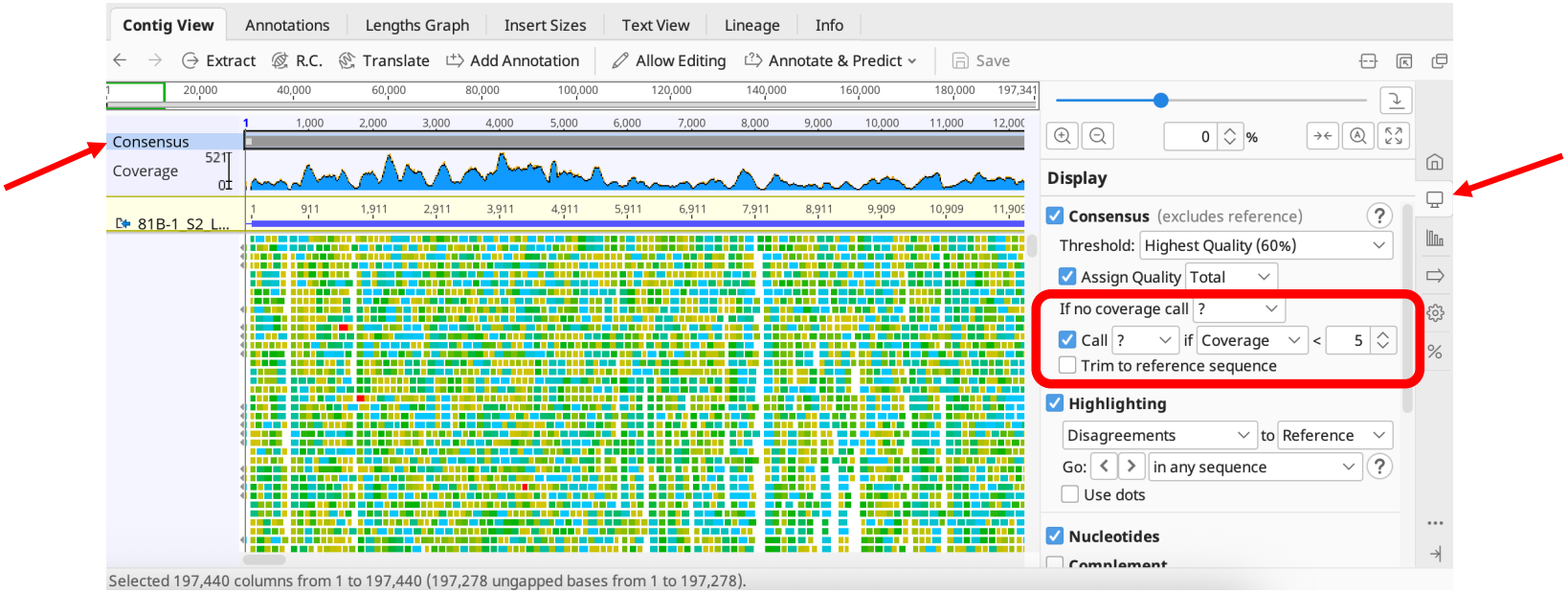


Aligned “reads” (.bam)



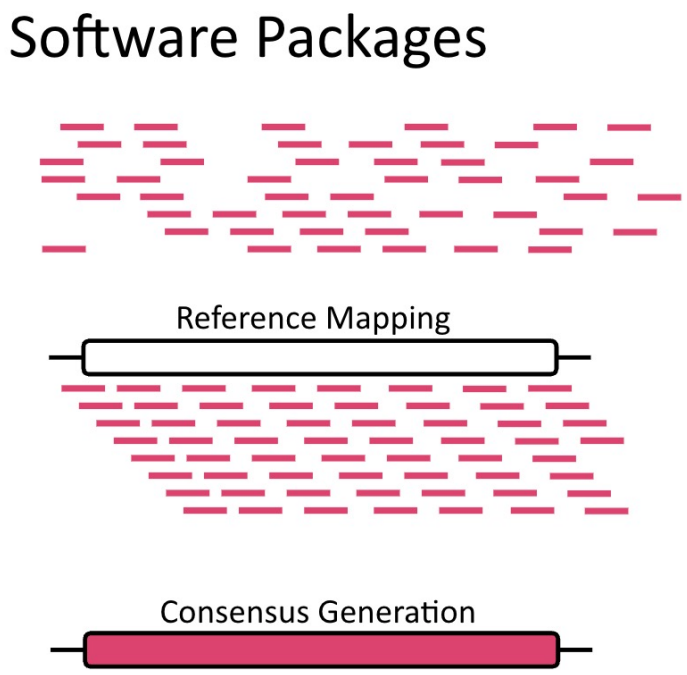
Consensus sequence (.fasta)

Generate a consensus sequence from aligned reads.



To generate a consensus sequence, just select “Consensus” at the top. Make sure you have correct parameters assigned. These can be modified in the “Display” tab on the right (looks like a computer screen). Make sure to select the Call if Coverage is < X reads. Remember depending on the type of sequencing data you are using (Illumina vs ONT) this number can vary. Additionally, it’s safer to call with ? than N because it doesn’t make the assumption there are nucleotides at positions with less than the desired coverage.

Remember there are multiple tools that can be used for these steps and sequence data.



geneious prime

GUI. Any data type.

ARTIC

CLI. MinION data.

https://github.com/artic-network/fieldbioinformatics

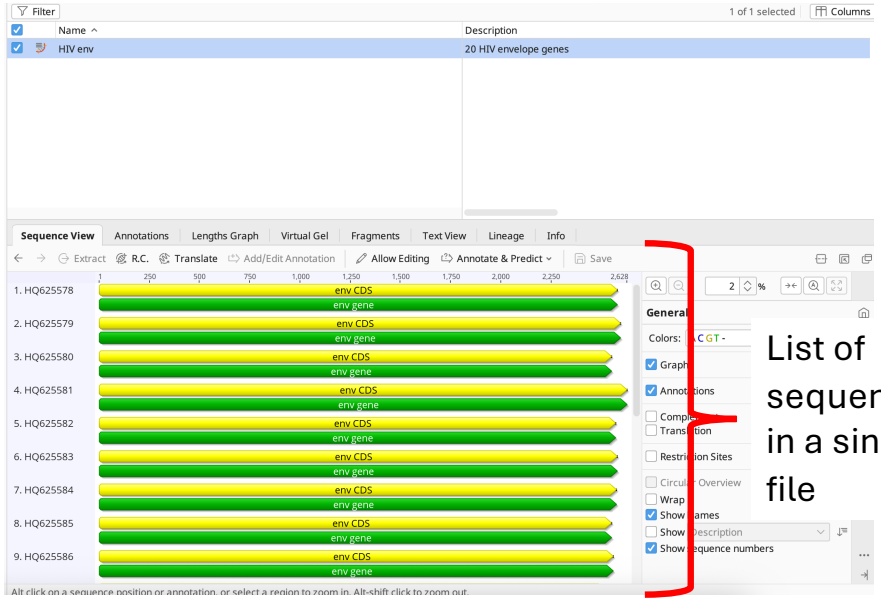
iVAR

CLI. Illumina data.

https://github.com/andersen-lab/ivar

Geneious can be useful to perform and visualize alignments.

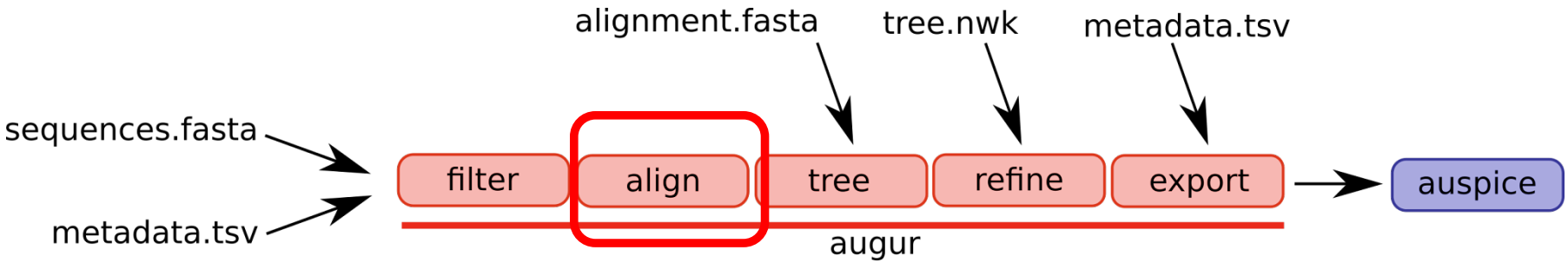
1. First, let's download Mafft (this is the aligner tool that Nextstrain uses). Similar to BBDuk, you'll download install it through Tools -> Plugins -> Mafft (Install). Note: No screenshot shown here to challenge you to you remember how to download tools! (If needed look back to slide 8)



List of sequences in a single file

2. Select the sequences that will be aligned. Note, sequences can be separate or they can be a sequence list. For example in this case, I am using the HIV env sequences provided by Geneious. These are all in one single file that is a list of sequences.

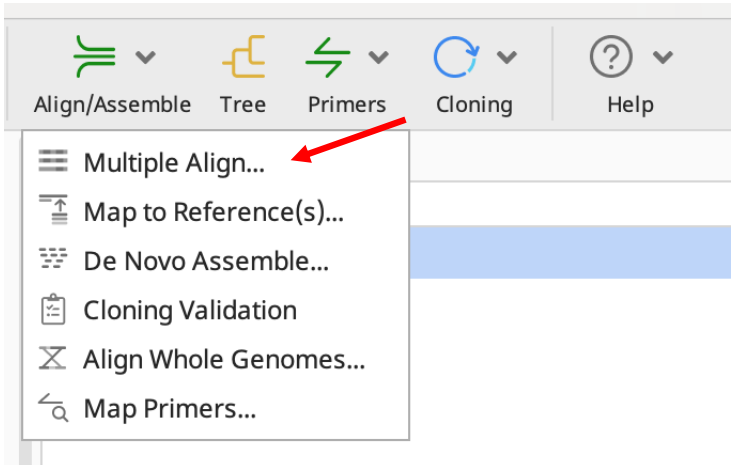
Even though Nextstrain will perform the alignment step in the workflow, this analysis can be done and visualized in Geneious.



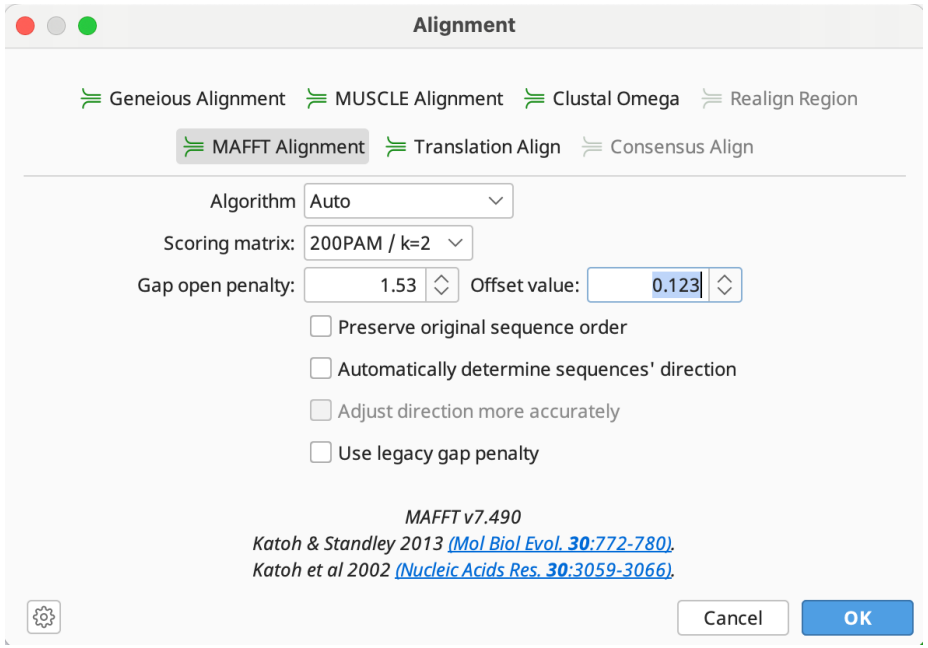

```
DENV1  AAAAGTCAGGTCAAACGCAGCTATTGGAGCAGTGTTCGTTGATGAAAATCA
DENV2  AAAGGTGAGAAGCAATGCAGCCTTGGGGGCCATATTCACTGATGAGAACAA
DENV3  AAAGGTCAGAACTAACGCAGCCATGGGCGCCGTTTTCACAGAGGAGAACCA
DENV4  AAAAGTTAGATCAAACGCAGCCATAGGCGCAGTCTTTCAGGAAGAACAGGG
ZIKV   CAAGGTGCGCAGCAATGCAGCACTGGGAGCAATATTTGAAGAGGAAAAAGA
WNV    AAAAGTCAACAGTAATGCCGCCCTAGGAGCGATGTTTGAAGAACAGAACCA
YFV    AAAAGTCCGAAGTCATGCAGCCATTGGAGCTTACCTGGAAGAACAAGAACA
POWV   GAAGGTGAGGTCCAACGCTGCTCTAGGTGCATGGTCGGATGAACAGAATAA

       *AA*GT********A*GC*GC**T*GG*GC**********GA**A**A***
```

Geneious can be useful to perform and visualize alignments.



3. With the sequences selected, go to Align/Assembly tool and then "Multiple Align"



4. Select MAFFT Alignment. You could play with the Algorithm used, but I typically use default settings since Auto should choose the best algorithm.

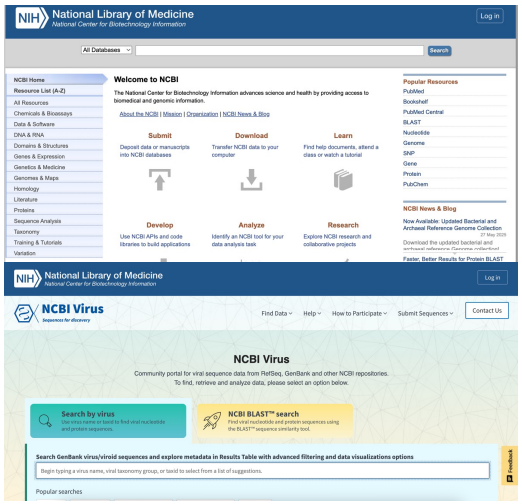
Geneious can be useful to perform and visualize alignments.

5. Now when you select the alignment file, you can visualize the alignment. In short, Green means perfectly conserved, Yellow means well conserved but variation (the height of the bars indicate the divergence), and Red means poor conservation. You can also see nucleotide statistics that indicate things such as total Pairwise Identity across the sequences.

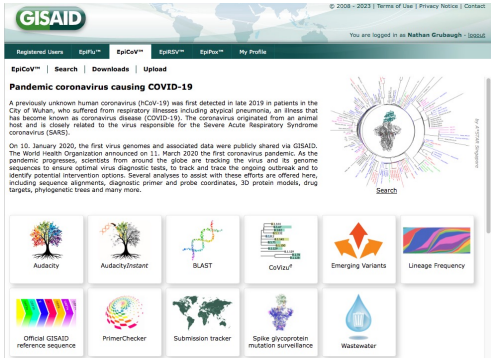
Geneious Exercises

Databases for accessing pathogen genomic datasets.

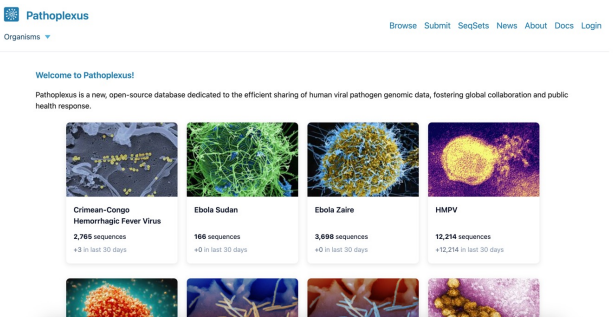
There are different databases available, but we will focus on 3 databases for this workshop.



NCBI (including NCBI Virus)



GISAID



Pathoplexus

Pros and Cons of Each Database

Feature	NCBI GenBank/SRA	GISAID	PathoPlexus
Main Focus	Broad genomic archive (all life forms)	Curated viral genomes (mainly respiratory pathogens)	Curated viral genomes
Organism Scope	All organisms (viruses, bacteria, eukaryotes)	Viruses only	Viruses only
Data Types	Raw reads (SRA), assemblies, annotations	Assembled viral genomes with metadata	Assembled viral genomes with curated, standardized metadata
Curation Level	Mixed: user-submitted + some curated	User submitted and manual QC	Highly curated; strict quality filters applied
Sequence Downloads	FASTA, GenBank, raw reads via SRA	FASTA (aligned/unaligned), metadata specific for Nextstrain formats	FASTA + tab-delimited metadata via website
Metadata Quality	Inconsistent; varies by submitter	Rich and mostly clean, but varies across submissions	Clean and standardized; curated for consistency
Lineage/Genotype Info	Often missing or scattered	Yes (e.g., Nextstrain, PANGO lineages)	Yes; built-in for many viral pathogens
Registration Required?	No, but benefits with registration	Yes (requires account and agreement)	No, but highly encouraged
Data Use Restrictions	Public domain	Restricted; must acknowledge GISAID use	Openly accessible with publication citation requested; some data restrictions
Access Interface	Web portal and command-line tools	Web portal	Web portal
Downsides	Messy metadata, variable quality	Access restrictions, no raw data	Limited pathogen scope (viruses only); no raw reads

Pathoplexus
@pathoplexus.org

I know that name!!!!

Thanks to huge efforts by Uganda's Central Public Health Laboratory, Uganda Ministry of Health, UVRI, Africa CDC, & SANBI, UWC the sequence from the Ebola Sudan case in Kampala is now available on Pathoplexus pathoplexus.org/seq/PP_0011C... w full analysis here: virological.org/t/near-real-...

February 5, 2025 at 5:52 PM Everybody can reply

Databases Exercise

Data acquisition and curation

- Truly, this is a difficult part of Nextstrain, acquiring the right data
- Furthermore, parsing the metadata and collecting it into a useable format requires fairly robust coding/bioinformatic tools
- For this workshop, we will rely on Pathoplexus and already maintained Nextstrain pathogens because they curated and maintain this data

Nextstrain ingest workflow

- Although we will focus on using already curated data, Nextstrain does offer great tutorials and coding structure to create an ingest workflow

- What is an ingest workflow? It is a workflow that acquires (ingests) “public data from NCBI and outputs curated metadata and sequences that can be used as input for the phylogenetic workflow”

What are the steps to an ingest workflow?

- **What is the resource you are using to acquire your data?** Nextstrain only has documentation for NCBI. This is ok, because it likely is the only source needed. However, the ingest workflow does have steps to check the uncurated metadata to make sure it seems sufficient.
- Note: Only viruses can be downloaded with NCBI taxonID; other organisms must make use of Entrez! (This isn't really clear in their documentation at the moment)

What are the components to an ingest workflow?

- Core components:
  - Snakefile: Automates all ingest steps.
  - config/: Taxon- or pathogen-specific settings (e.g., NCBI taxon ID).
  - rules/: Modular Snakemake rules for the automation Snakefile
  - scripts/: Python scripts to parse and format metadata.
  - data/: Downloaded data and intermediate files
  - logs/: output logs for each step
  - results/: Final output files: sequences.fasta, metadata.tsv.

What are the basic steps to a ingest workflow?

- Fetch the data:
  - Uses the taxonID to download the sequences and metadata
  - Important: Only viruses can be downloaded with taxonID. Otherwise, use the entrez ID
- Curate the metadata that was downloaded:
  - Rename the fields
  - Verify strain names
  - Format dates to YYYY-MM-DD and mask missing dates
  - Transfer geographic location identifiers
  - Capitalize all first letters for standardization
  - Capture author names
  - Additional user-specific configurations (geolocation rules/wildcards, specific record annotations, etc.)
- Nextstrain is working to integrate Nextclade, where available

Nextstrain Ingest Examples

- A more specific overview of each of these steps can be found here: https://docs.nextstrain.org/en/latest/tutorials/creating-a-pathogen-repo/creating-an-ingest-workflow.html
- The generalized ingest workflow and guide can be found here: https://github.com/nextstrain/pathogen-repo-guide
- Mpox example: https://github.com/nextstrain/mpox/tree/master/ingest
- Zika example: https://github.com/nextstrain/zika/tree/main/ingest
- SARS-CoV2 example (really complex as it pulls in GISAID too): https://github.com/nextstrain/ncov-ingest

Questions?