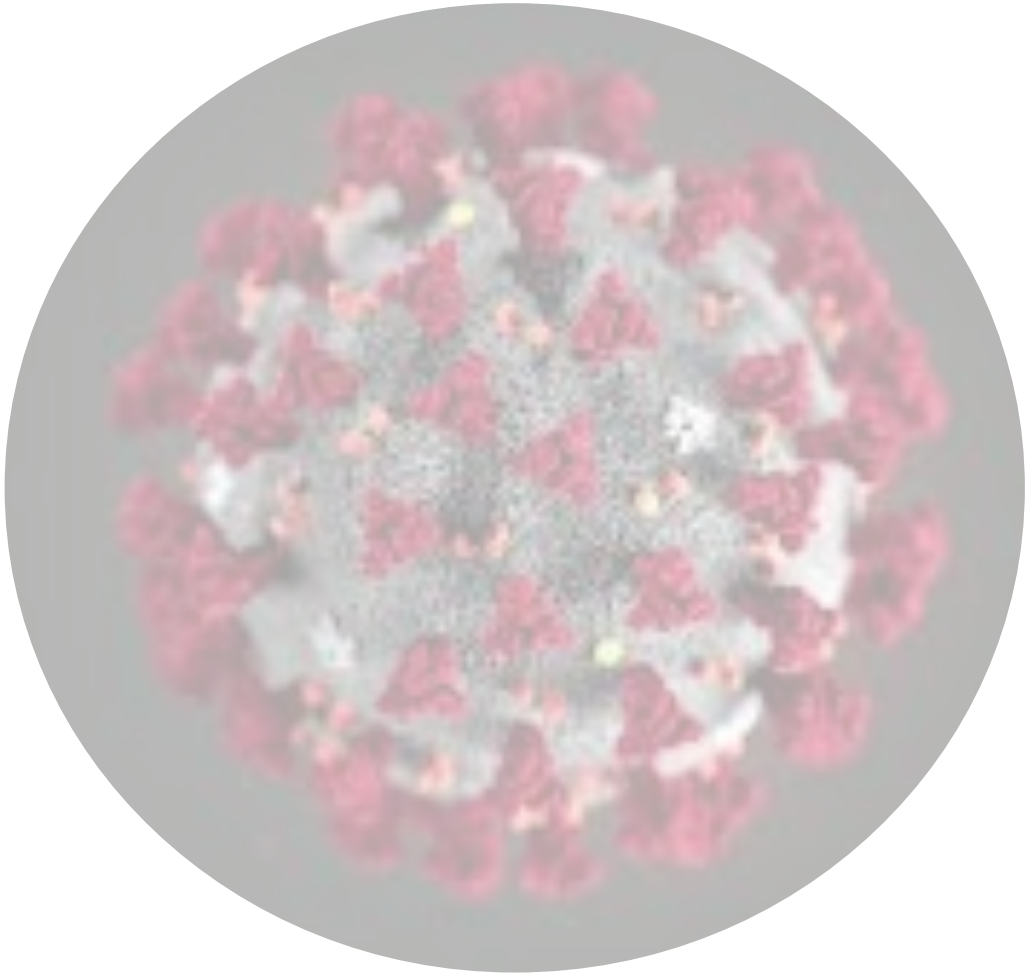
Understanding Modern Genomic Epidemiology Lecture 2



JOSEPH FAUVER, PH.D.

ASSISTANT PROFESSOR

UNMC COPH DEPARTMENT OF EPIDEMIOLOGY

5/7/2025

UNIVERSITY OF NEBRASKA MEDICAL CENTER℠
COLLEGE OF PUBLIC HEALTH

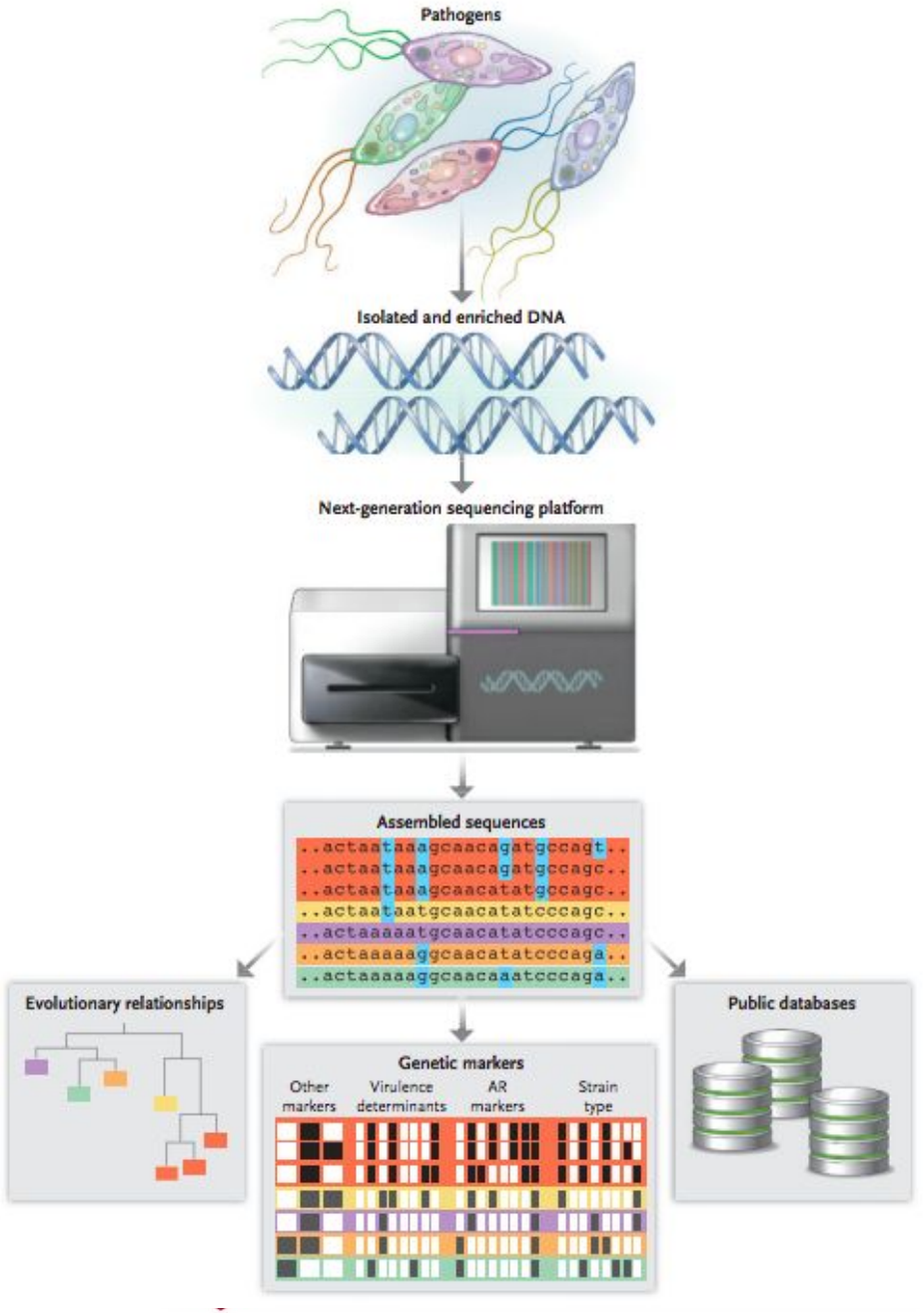
What does this look like?

"Wet" lab:

- Clinical samples/Microbial isolation
- RNA or DNA extraction
- Library preparation
- Next-Generation Sequencing

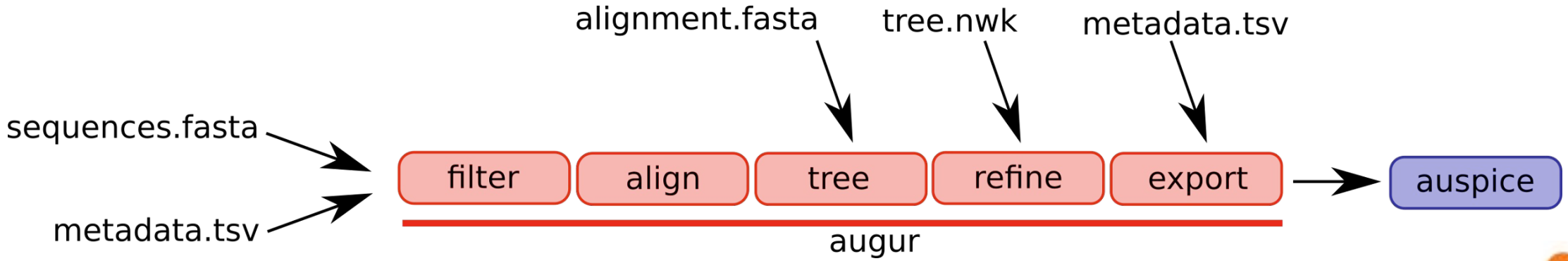
"Dry" lab:

- Data processing
- Primary genomic analyses
- Data interpretation
- Submission to repositories



What is next?

- **NGS data, consensus sequence generation, data repositories**
- Description of Nextstrain (what it is, what it isn't, how it works, etc.)
- Basics of phylogenetic tree interpretations



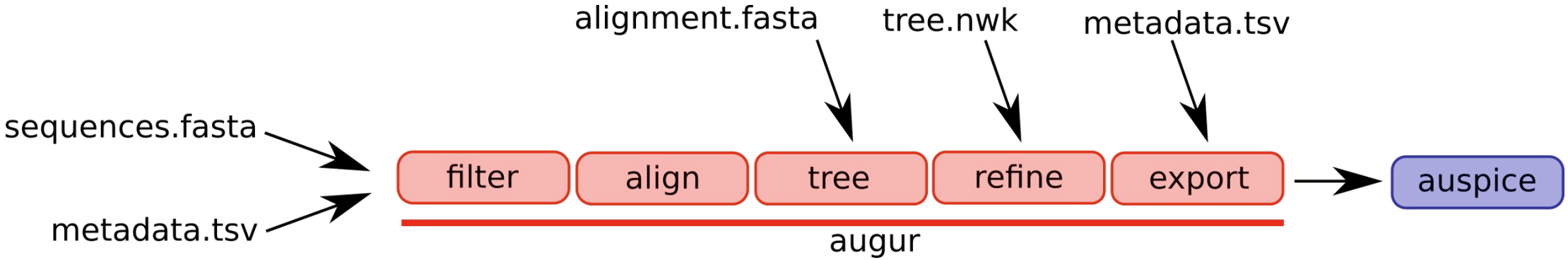
Nextstrain

Real-time tracking of pathogen evolution

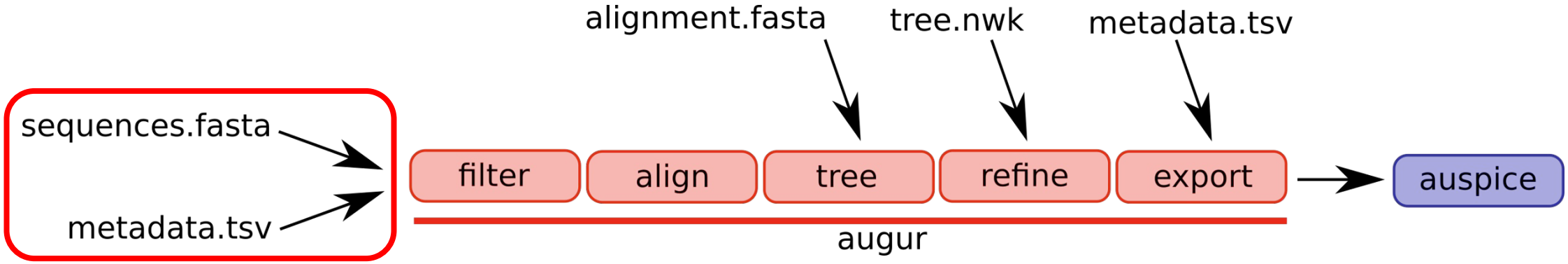
geneious prime

UNIVERSITY OF NEBRASKA MEDICAL CENTER
COLLEGE OF PUBLIC HEALTH

Preparing to run a Nextstrain build

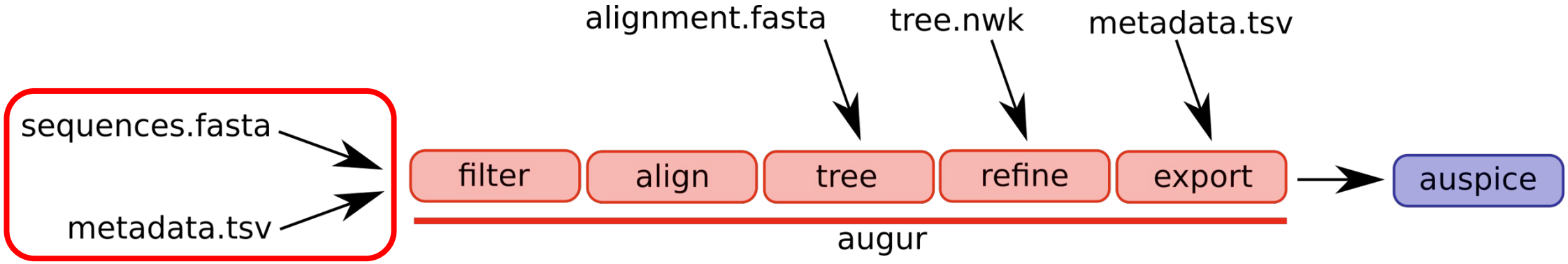


Preparing to run a Nextstrain build



- **Next-Generation Sequencing Data -> Consensus Sequence Generation**
  - Data types (.fastq, .fasta, .bam)
  - Reference mapping
  - Variant calling
- **Identifying relevant data sources/repositories for contextual data**
  - Pathoplexus
  - NCBI GenBank (SRA, Virus, Microbe)
  - GISAID
  - PlsmoDB
  - BacWGSTdb

Preparing to run a Nextstrain build

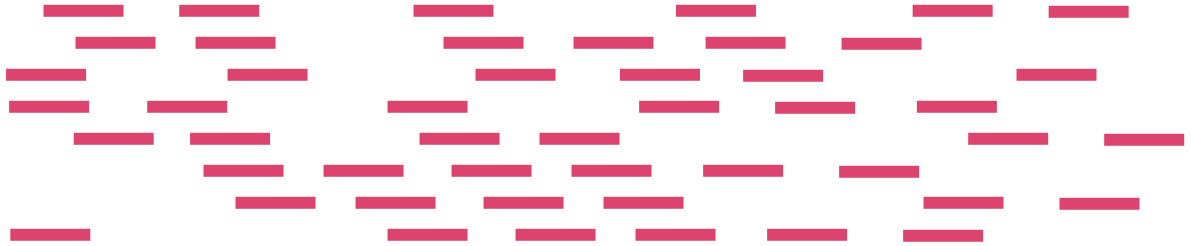


- **Next-Generation Sequencing Data -> Consensus Sequence Generation**
  - Data types (.fastq, .fasta, .bam)
  - Reference mapping
  - Variant calling

geneious prime

- **Identifying relevant data sources/repositories for contextual data**
  - Pathoplexus
  - NCBI GenBank (SRA, Virus, Microbe)
  - GISAID
  - PlsmoDB
  - BacWGSTdb

NGS data -> Consensus sequence

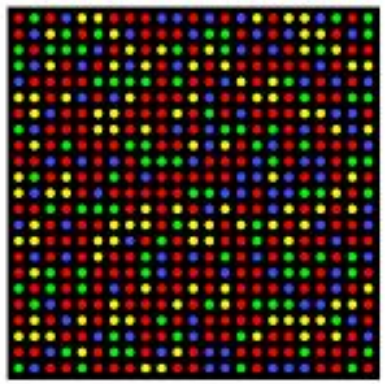


Raw “reads” (.fastq)

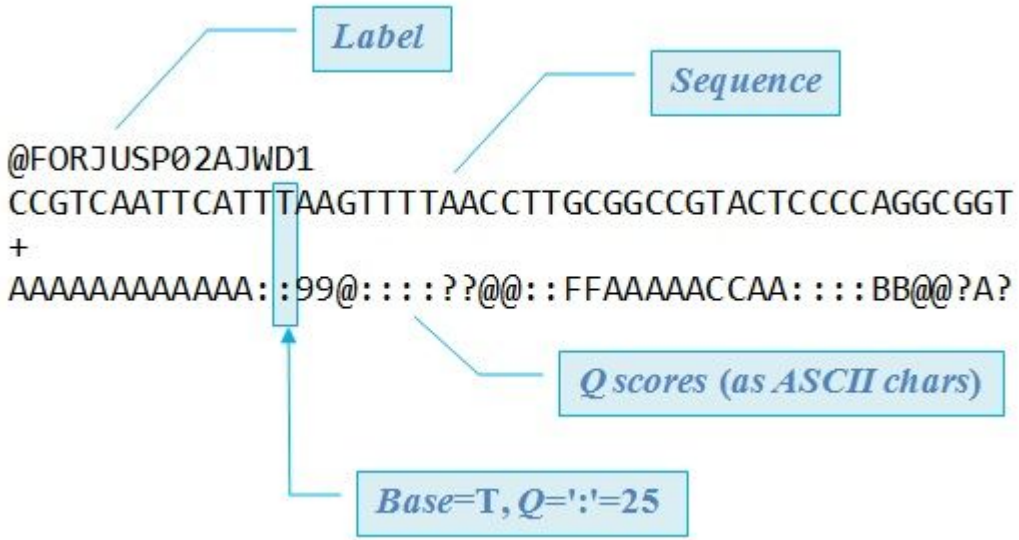
NGS Data Formats

Nanopore data

Illumina data

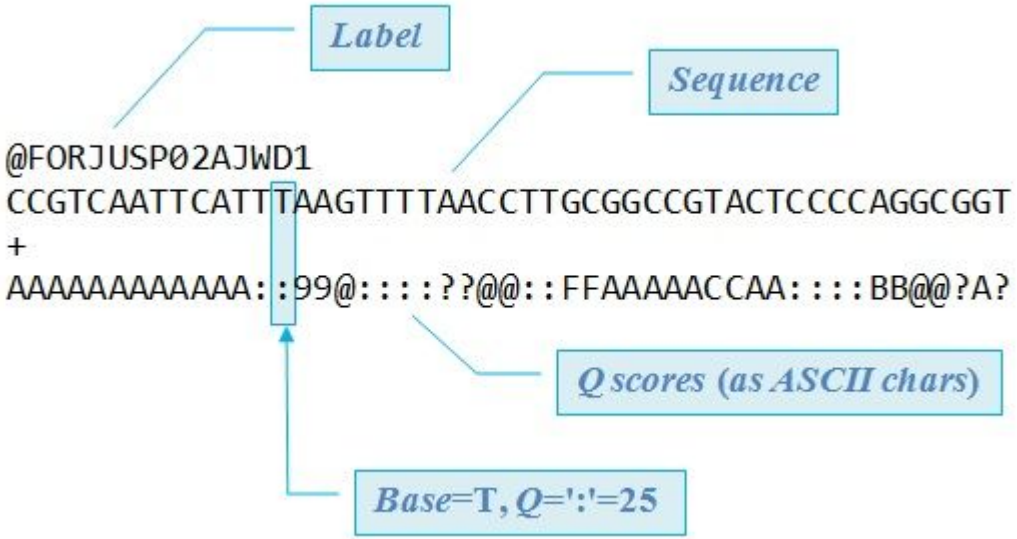


.fastq



Anatomy of a .Fastq vs .Fasta file

.fast**q** file



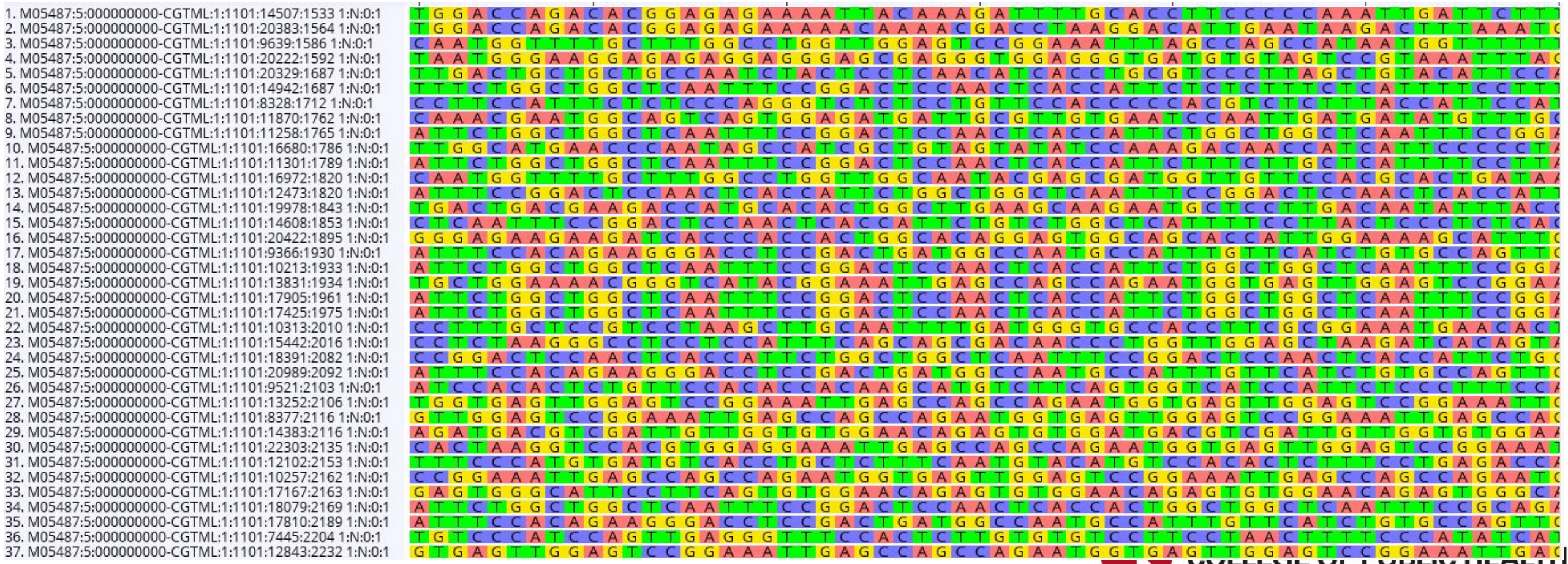
.fasta file

```
>NG_008679.1:5001-38170 Homo sapiens paired box 6 (PAX6)
ACCCTCTTTTCTTATCATTGACATTTAAACTCTGGGGCAGGTCCTCGCGTAGAACGCGGCTGTCAGATCT
GCCACTTCCCCTGCCGAGCGGCGGTGAGAAGTGTGGGAACCGGCGCTGCCAGGCTCACCTGCCTCCCCGC
CCTCCGCTCCCAGGTAACCGCCCGGGCTCCGGCCCCGGCCCGGCTCGGGGCCCGCGGGGCCTCTCCGCTG
CCAGCGACTGCTGTCCCCAAATCAAAGCCCGCCCCAAGTGGCCCCGGGGCTTGATTTTTGCTTTTAAAAG
GAGGCATACAAAGATGGAAGCGAGTTACTGAGGGAGGGATAGGAAGGGGGGTGGAGGAGGGACTTGTCTT
TGCCGAGTGTGCTCTTCTGCAAAAGTAGCAAAATGTTCCACTCCTAAGAGTGGACTTCCAGTCCGGCCCT
GAGCTGGGAGTAGGGGGCGGGAGTCTGCTGCTGCTGTCTGCTAAAGCCACTCGCGACCGCGAAAAATGCA
GGAGGTGGGGACGCACTTTGCATCCAGACCTCCTCTGCATCGCAGTTCACGACATCCACGCTTGGGAAAG
TCCGTACCCGCGCCTGGAGCGCTTAAAGACACCCTGCCGCGGGTCGGGCGAGGTGCAGCAGAAGTTTCCC
GCGGTTGCAAAGTGCAGATGGCTGGACCGCAACAAAGTCTAGAGATGGGGTTCGTTTCTCAGAAAGACGC
```

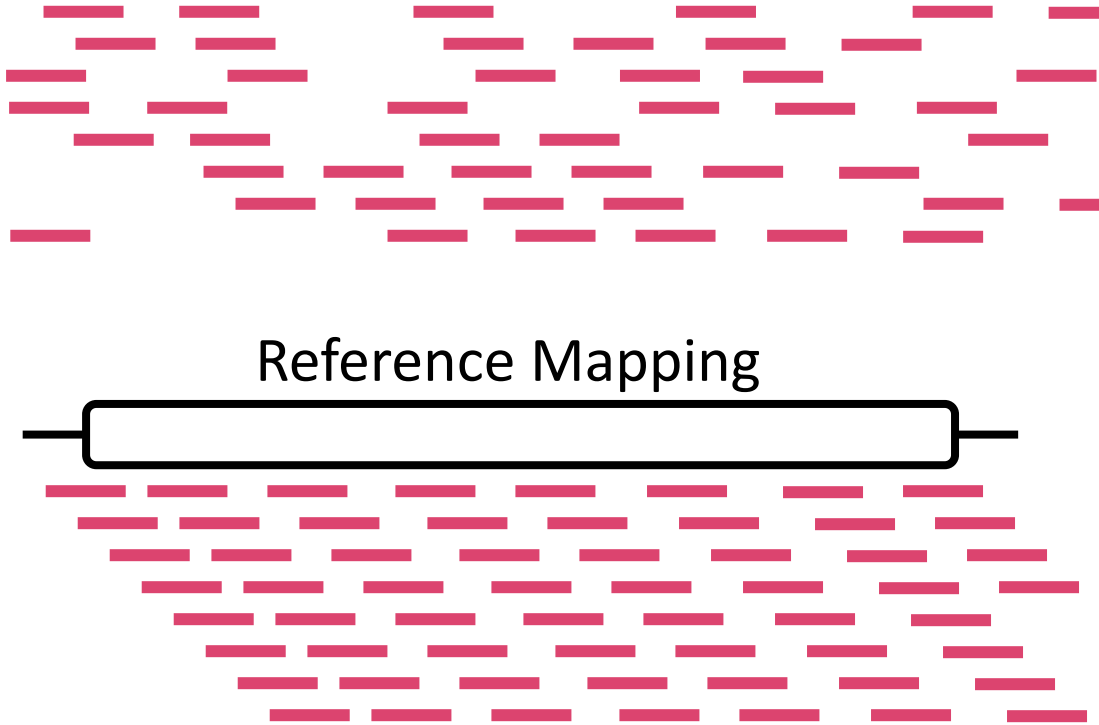
Sequencing reads in Geneious

Header

Bases



NGS data -> Consensus sequence



Raw “reads” (.fastq)

Aligned “reads” (.bam)

Steps to align reads to reference genome

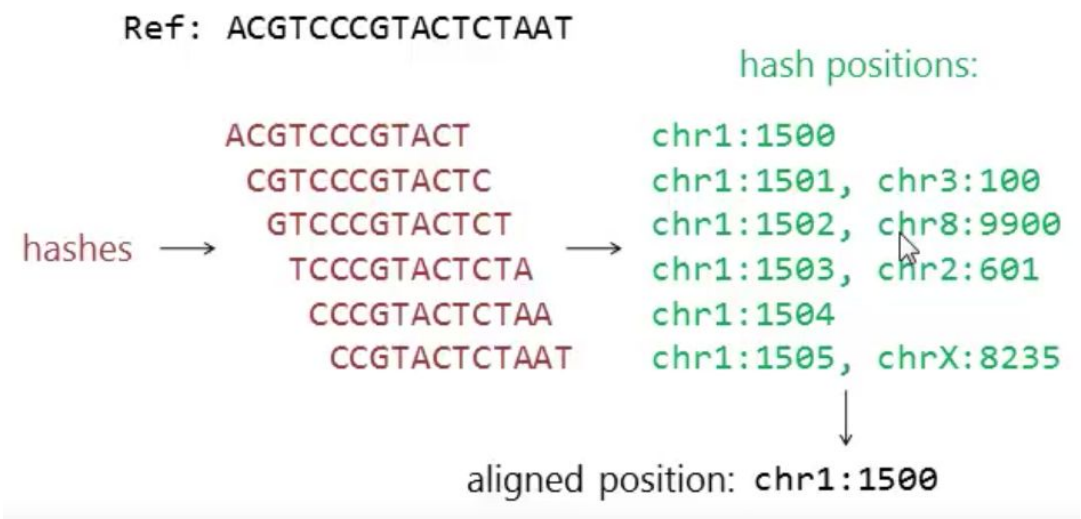
1. Quality trimming - “Soft clip”

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

2. Search for best match on given reference genome with the following considerations:
   - Score of match
   - Mismatch penalty
   - Gap penalty

Can also do competitive alignments with multiple reference genomes

Algorithms: Hashing



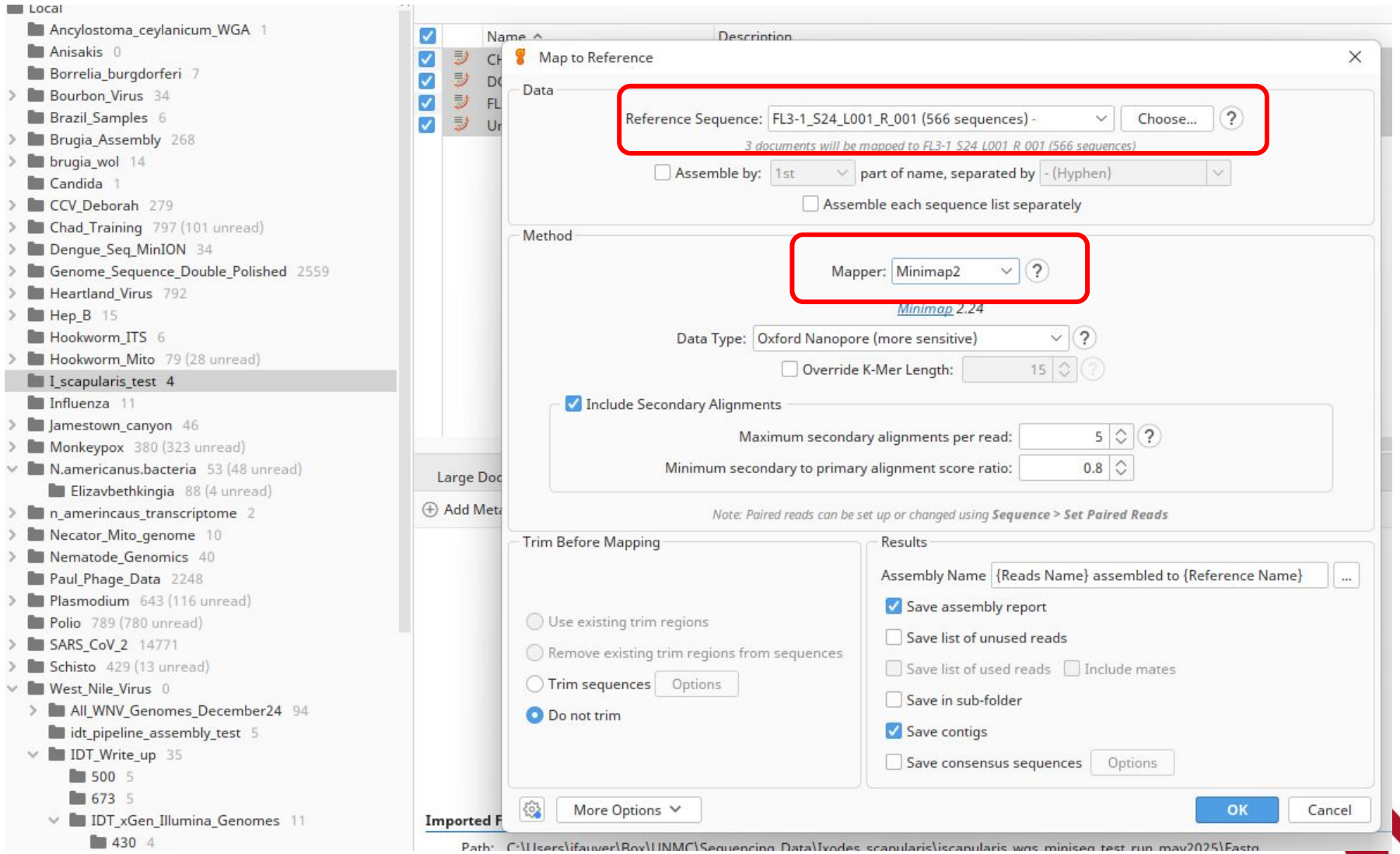
Output format: SAM or BAM (binary) files

- SAM = Sequence Alignment Map

What software to choose?

- Burrows-Wheeler Aligner (BWA)
  - Recommended for Illumina data
  - Newer version = BWA-MEM recommended
  - http://bio-bwa.sourceforge.net/
- Mimimap2
  - Recommended for Nanopore data
  - https://lh3.github.io/minimap2/minimap2.html
- Many other possibilities
- Can be done in Geneious!

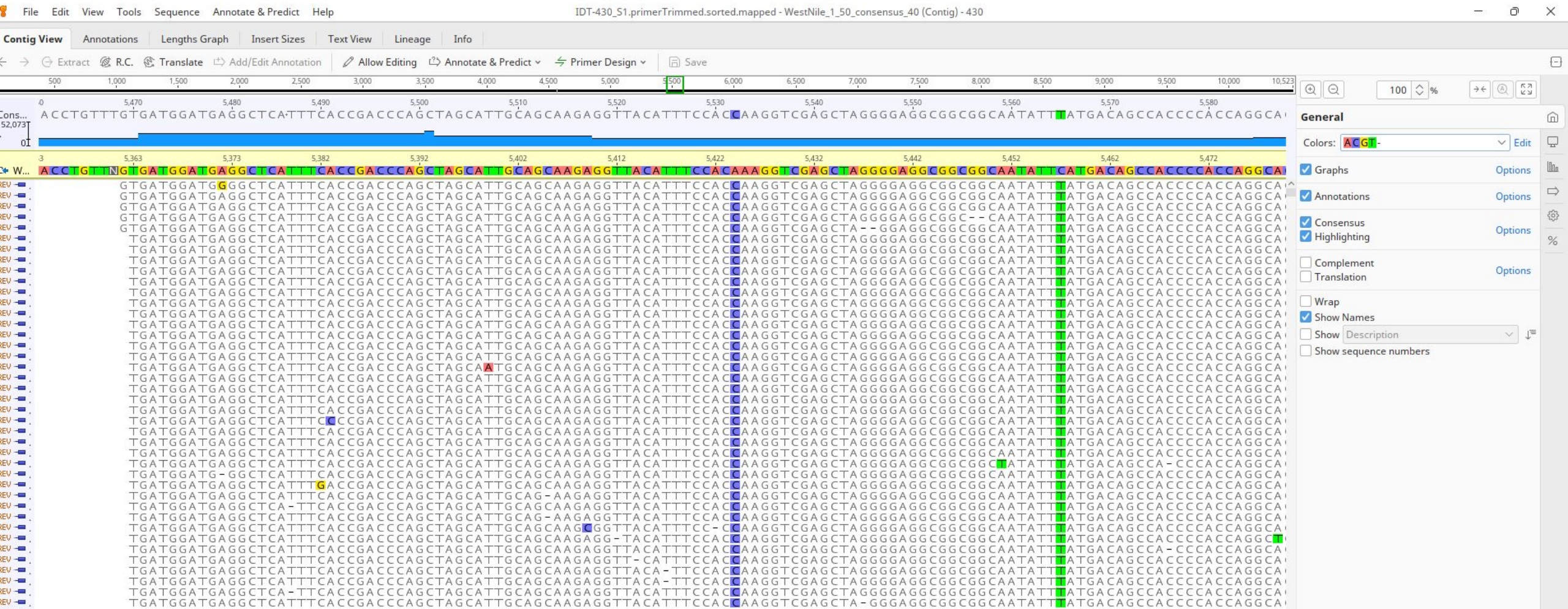
Generating a .bam file



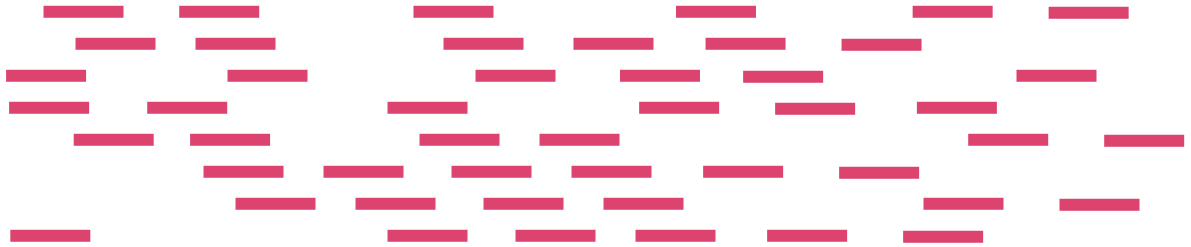
- Select a reference sequence that is most similar to your target of interest

- Sequence Length
- Sequence Content

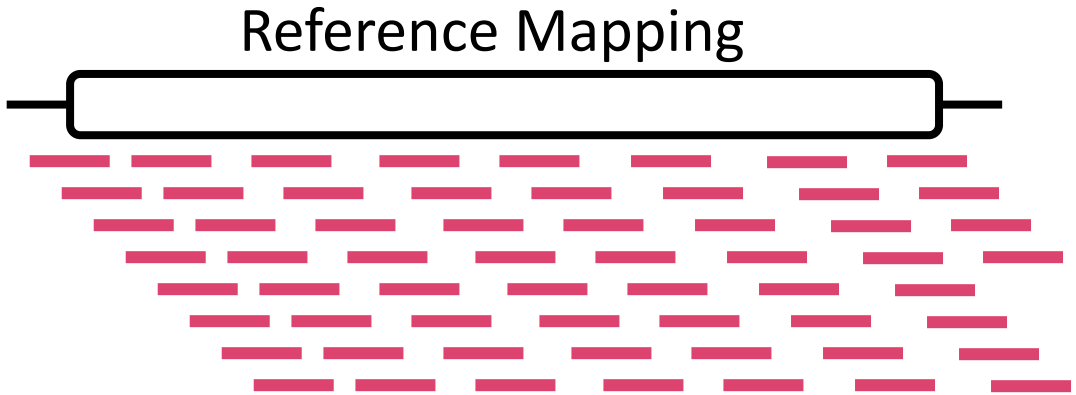
Generating a .bam file



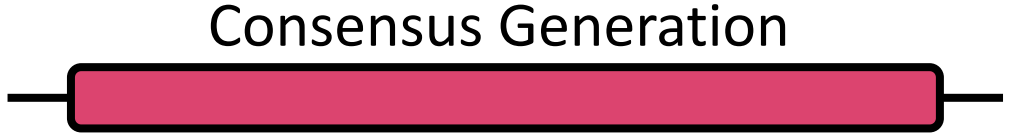
NGS data -> Consensus sequence



Raw “reads” (.fastq)



Aligned “reads” (.bam)



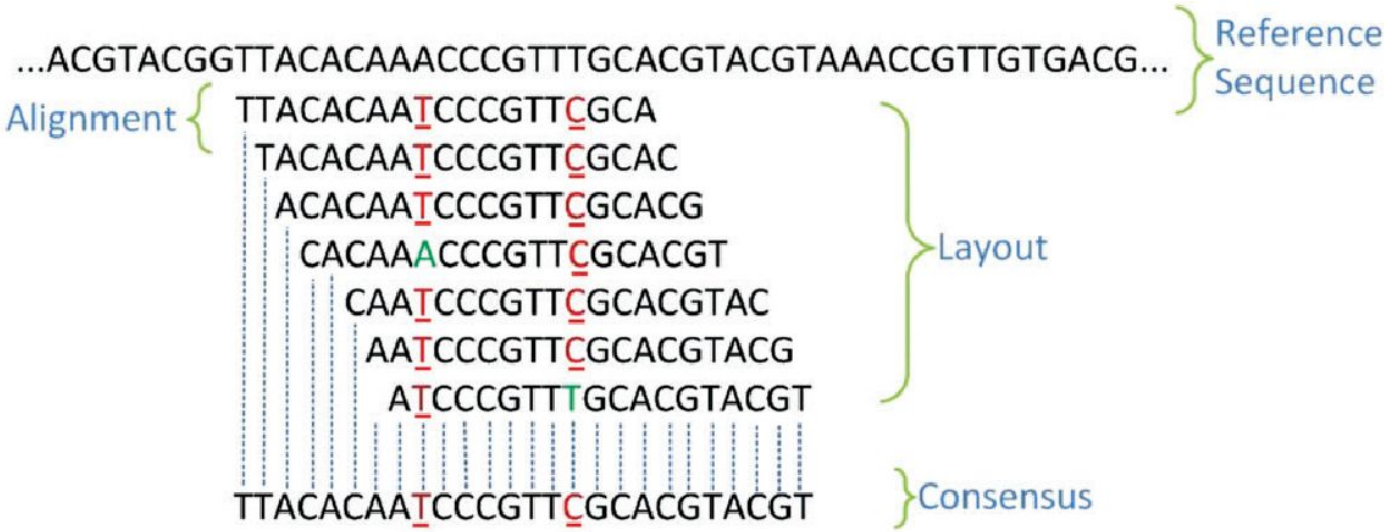
Consensus sequence (.fasta)

Consensus sequence genomes

- Primary input into any phylogenetic analysis pipeline (e.g. Nextstrain)
- Represents the *most common* base at any position in the genome of interest
- .bam files are used to call consensus sequences, which are represented in .fasta files

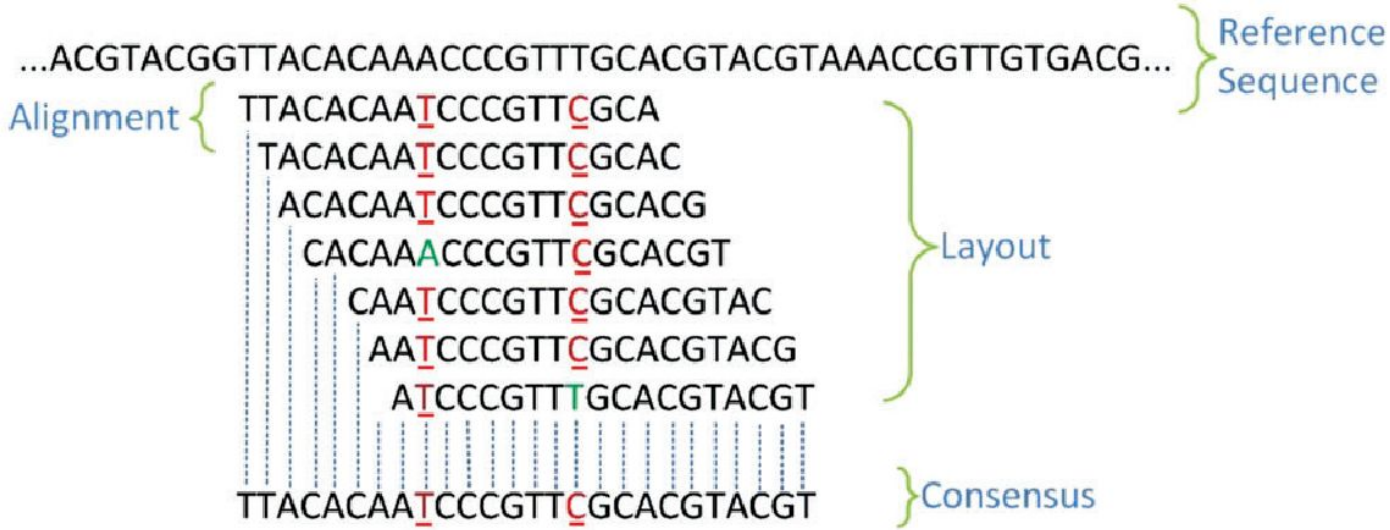
Consensus sequence genomes

- Primary input into any phylogenetic analysis pipeline (e.g. Nextstrain)
- Represents the *most common* base at any position in the genome of interest
- .bam files are used to call consensus sequences, which are represented in .fasta files



Considerations for Consensus Sequence

- Read depth- how much data do you need at a site to confidently call a base? Typically 5-20x
- Dealing with heterogeneity
  - Majority of bases will be reference sequence
  - SNPS (Single Nucleotide Polymorphisms) are often clearly represented
  - Set thresholds to call ambiguous nucleotides



Considerations for Consensus Sequence

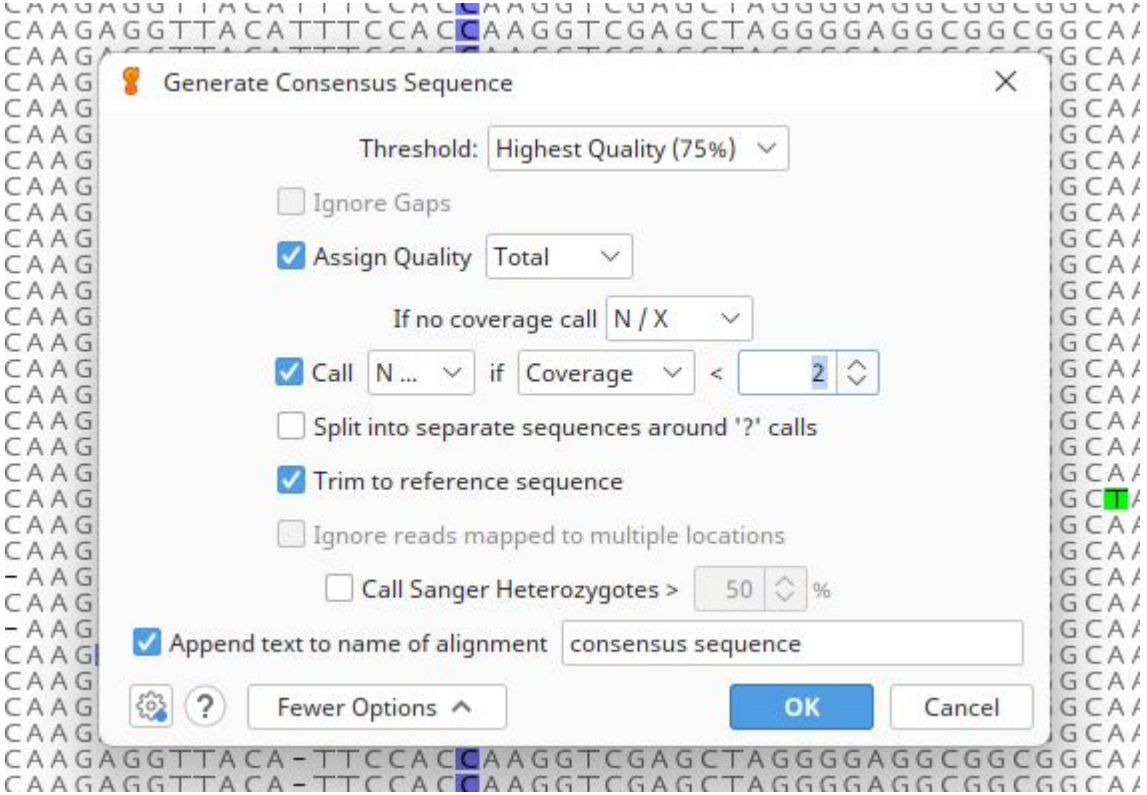
- Read depth- how much data do you need at a site to confidently call a base? Typically 5-20x
- Dealing with heterogeneity
  - Majority of bases will be reference sequence
  - SNPS (Single Nucleotide Polymorphisms) are often clearly represented
  - Set thresholds to call ambiguous nucleotides

When in doubt, use “N”, not the reference sequence!

Symbol[12]	Description	Bases represented				
A	adenine	A				1
C	cytosine		C			
G	guanine			G		
T	thymine				T	
U	uracil				U	
W	weak	A			T	2
S	strong		C	G		
M	amino	A	C			
K	keto			G	T	
R	purine	A		G		
Y	pyrimidine		C		T	
B	not A (**B** comes after A)		C	G	T	3
D	not C (**D** comes after C)	A		G	T	
H	not G (**H** comes after G)	A	C		T	
V	not T (**V** comes after T and U)	A	C	G		
N	any base (not a gap)	A	C	G	T	4

Considerations for Consensus Sequence

- Read depth- how much data do you need at a site to confidently call a base? Typically 5-20x
- Dealing with heterogeneity
  - Majority of bases will be reference sequence
  - SNPS (Single Nucleotide Polymorphisms) are often clearly represented
  - Set thresholds to call ambiguous nucleotides

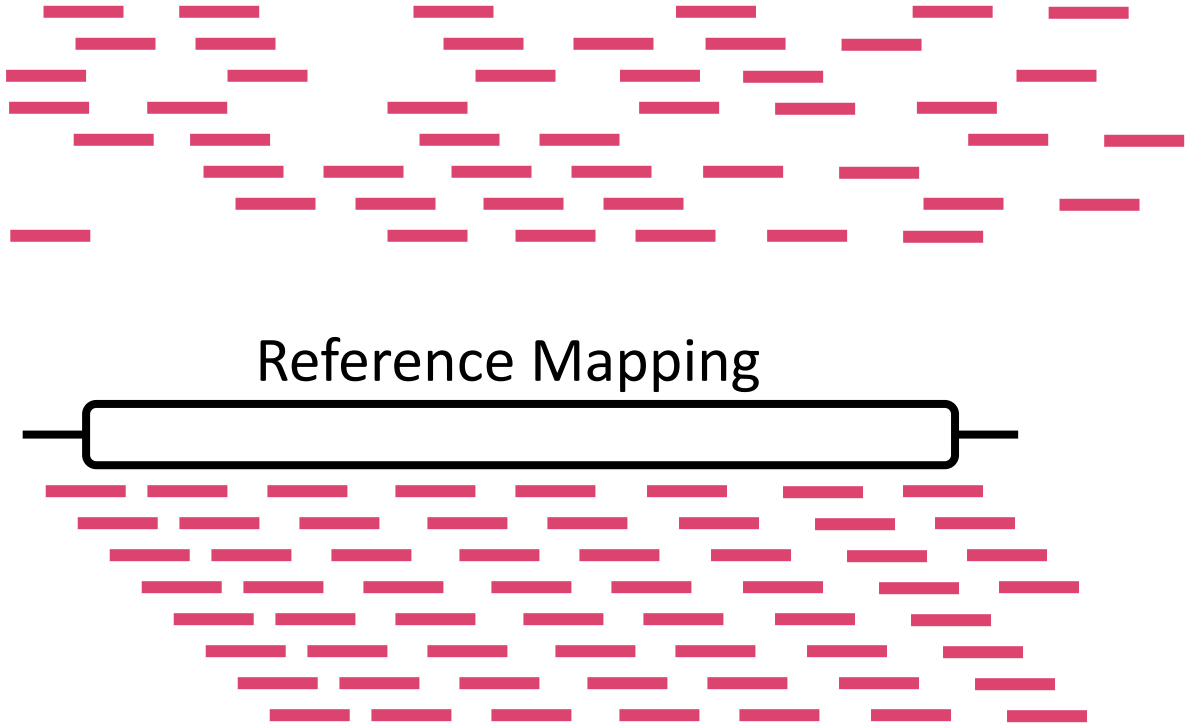


When in doubt, use “N”, not the reference sequence!

geneious prime

UNIVERSITY OF NEBRASKA MEDICAL CENTER
COLLEGE OF PUBLIC HEALTH

Software Packages



geneious prime

GUI. Any data type.

ARTIC

CLI. MinION data.

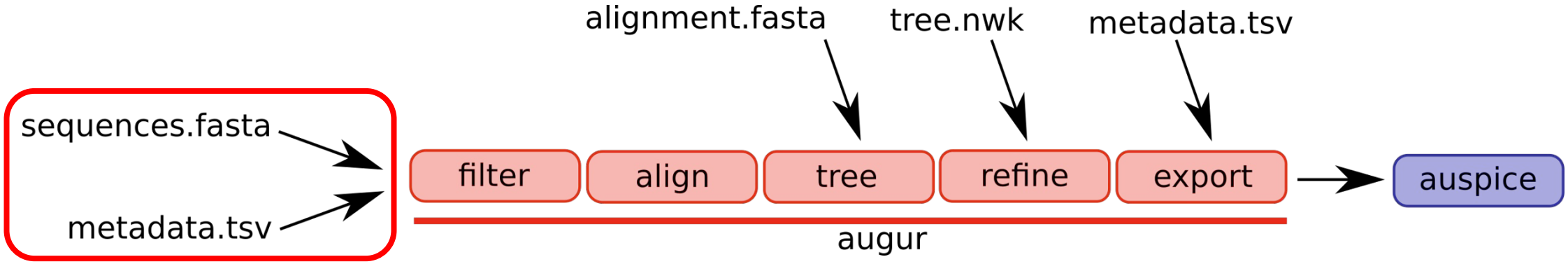
https://github.com/artic-network/fieldbioinformatics

iVAR

CLI. Illumina data.

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

Preparing to run a Nextstrain build



- **Next-Generation Sequencing Data -> Consensus Sequence Generation**
  - Data types (.fastq, .fasta, .bam)
  - Reference mapping
  - Variant calling
- **Identifying relevant data sources/repositories for contextual data**
  - Pathoplexus
  - NCBI GenBank (SRA, Virus, Microbe)
  - GISAID
  - PlsmoDB
  - BacWGSTdb

Identifying relevant data repositories

- NCBI GenBank (fasta) https://www.ncbi.nlm.nih.gov/genbank/
  - NCBI SRA (fastq or bam) https://www.ncbi.nlm.nih.gov/sra
  - NCVI Virus: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/
  - NCBI Microbes: https://www.ncbi.nlm.nih.gov/genome/microbes/
- European Nucleotide Archive: https://www.ebi.ac.uk/ena/browser/home
  - EnsemblBacteria: https://bacteria.ensembl.org/index.html
- DNA DataBank of Japan: https://www.ddbj.nig.ac.jp/index-e.html
- Other specialty
  - GISAID (viruses): https://www.gisaid.org/
  - BacWGSTdb (bacteria): http://bacdb.cn/BacWGSTdb/
  - PlasmoDB (plasmodium): https://plasmodb.org/plasmo/app/
  - Pathoplexus (handful of pathogens): https://pathoplexus.org/

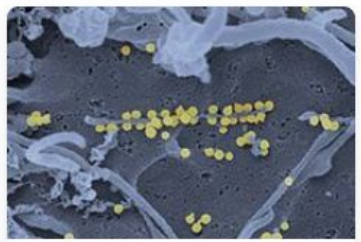
Pathoplexus

Organisms

Browse Submit SeqSets News About Docs Login

Welcome to Pathoplexus!

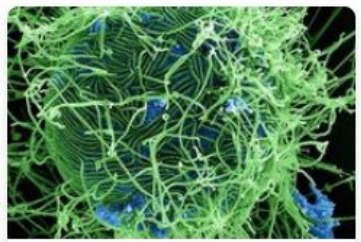
Pathoplexus is a new, open-source database dedicated to the efficient sharing of human viral pathogen genomic data, fostering global collaboration and public health response.



Crimean-Congo Hemorrhagic Fever Virus

2,762 sequences

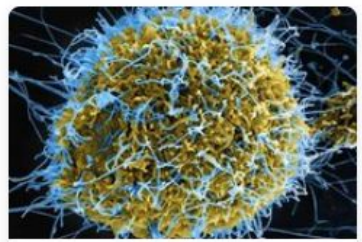
+0 in last 30 days



Ebola Sudan

166 sequences

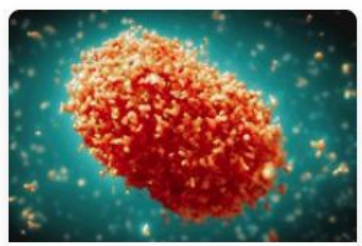
+0 in last 30 days



Ebola Zaire

3,698 sequences

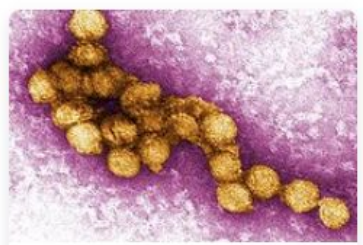
+0 in last 30 days



Mpox Virus

10,328 sequences

+88 in last 30 days



West Nile Virus

7,830 sequences

+3 in last 30 days

ncbi.nlm.nih.gov/labs/virus/vssi/#/

This is an NCBI Labs Experiment. Learn more.

National Library of Medicine
National Center for Biotechnology Information

nathan.grubaugh@y...

NCBI Virus
Sequences for discovery

About Us | Find Data | Help | How to Participate | Submit Sequences | Contact Us

Search by sequence
Use the NCBI BLAST™ tool to find similar viral nucleotide and protein sequences.

Search by virus
Use virus name or taxid to find viral nucleotide and protein sequences.

Search by virus name or taxonomy

Begin typing a virus name, viral taxonomy group, or taxid to select from a list of suggestions.

virus name or taxid

Popular searches

All viruses | Human viruses | Bacteriophages | New sequences (past one month) | **Up-to-date SARS-CoV-2**

GISAID

You are logged in as **Nathan Grubaugh** - logout

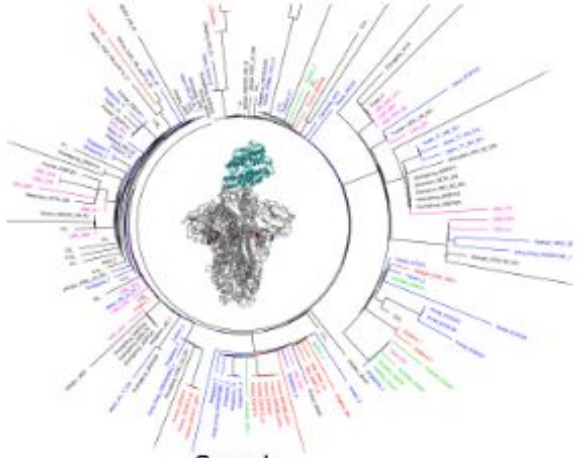
Registered Users | EpiFlu™ | EpiCoV™ | EpiRSV™ | EpiPox™ | My Profile

EpiCoV™ | **Search** | **Downloads** | **Upload**

Pandemic coronavirus causing COVID-19

A previously unknown human coronavirus (hCoV-19) was first detected in late 2019 in patients in the City of Wuhan, who suffered from respiratory illnesses including atypical pneumonia, an illness that has become known as coronavirus disease (COVID-19). The coronavirus originated from an animal host and is closely related to the virus responsible for the Severe Acute Respiratory Syndrome coronavirus (SARS).

On 10. January 2020, the first virus genomes and associated data were publicly shared via GISAID. The World Health Organization announced on 11. March 2020 the first coronavirus pandemic. As the pandemic progresses, scientists from around the globe are tracking the virus and its genome sequences to ensure optimal virus diagnostic tests, to track and trace the ongoing outbreak and to identify potential intervention options. Several analyses to assist with these efforts are offered here, including sequence alignments, diagnostic primer and probe coordinates, 3D protein models, drug targets, phylogenetic trees and many more.



Search

Audacity

AudacityInstant

BLAST

CoVizu[e]

Emerging Variants

Lineage Frequency

Official GISAID reference sequence

PrimerChecker

Submission tracker

Spike glycoprotein mutation surveillance

Wastewater

We need metadata in addition to genome data!

- Minimum information needed
  - Date of sample collection and/or onset of patient symptoms
    - YYYY-MM-DD; Only the year is not sufficient!
  - Location of sample collection and/or location of exposure
    - Country level minimum; often need higher resolution (beware of IRB restrictions)
  - Host and sample type used to generate the sequence
    - E.g., human serum, tissue, environmental, vector, etc;
- Additional information
  - Patient: Symptoms, outcomes, treatment/vaccination, PCR test results
  - Passage history
  - Sample collection methods
  - Sequencing & data processing methods (access to raw data)
- Can only release patient information that is approved by the IRB

We need metadata in addition to genome data!

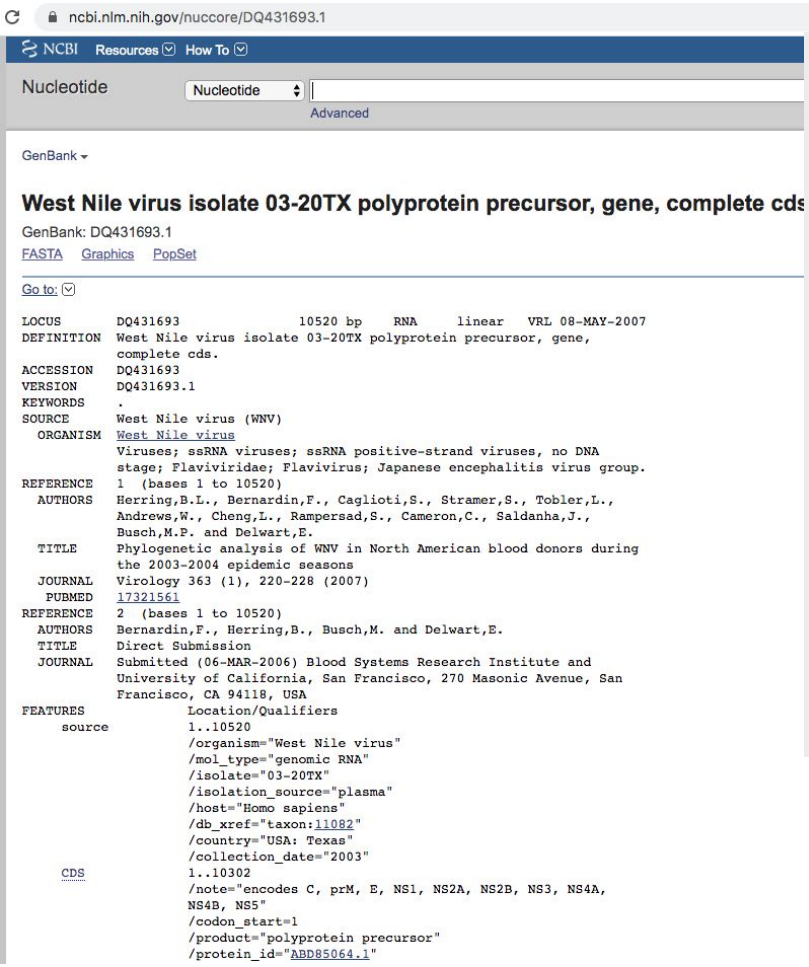
	A	B	C	D	E	F	G	H	K	R	S
1	strain	virus	accessior	date	region	country	state	division	segment	latitude	longitude
2	AF196835	wnv	AF196835	1999-XX-X	North Ame	USA	NY	NY/Bronx	genome	42.1657	-74.9481
3	AF206518	wnv	AF206518	1999-XX-X	North Ame	USA	CT	CT/Fairfiel	genome	41.5978	-72.7554
4	DQ211652	wnv	DQ211652	1999-XX-X	North Ame	USA			genome		
5	EF530047	wnv	EF530047	2000-XX-X	North Ame	USA			genome		
6	EF571854	wnv	EF571854	1999-XX-X	North Ame	USA			genome		
7	EF657887	wnv	EF657887	2000-XX-X	North Ame	USA			genome		
8	FJ527738	wnv	FJ527738	2001-XX-X	North Ame	USA	LA	LA	genome	31.1695	-91.8678
9	GQ379156	wnv	GQ379156	2001-07-X	North Ame	USA			genome		
10	GQ379157	wnv	GQ379157	2008-08-X	North Ame	USA	CA	CA	genome	36.1162	-119.682
11	GQ379158	wnv	GQ379158	2007-08-X	North Ame	USA	CA	CA	genome	36.1162	-119.682
12	GQ379159	wnv	GQ379159	2008-08-X	North Ame	USA	CA	CA	genome	36.1162	-119.682
13	GQ379161	wnv	GQ379161	2006-02-X	North Ame	USA			genome		
14	HM538578	wnv	HM538578	2001-XX-X	North Ame	USA	CT	CT	genome	41.5978	-72.7554
15	HM538579	wnv	HM538579	2003-XX-X	North Ame	USA	CT	CT	genome	41.5978	-72.7554
16	HM538580	wnv	HM538580	2005-XX-X	North Ame	USA	CT	CT	genome	41.5978	-72.7554
17	HM538581	wnv	HM538581	2007-XX-X	North Ame	USA	CT	CT	genome	41.5978	-72.7554
18	HM538582	wnv	HM538582	2004-XX-X	North Ame	USA	IL	IL	genome	40.3495	-88.9861
19	HM538583	wnv	HM538583	2004-XX-X	North Ame	USA	IL	IL	genome	40.3495	-88.9861
20	HM756655	wnv	HM756655	2003-XX-X	North Ame	USA	CT	CT	genome	41.5978	-72.7554
21	HM756674	wnv	HM756674	2004-XX-X	North Ame	USA	NY	NY	genome	42.1657	-74.9481
22	HQ596519	wnv	HQ596519	1999-XX-X	North Ame	USA			genome		
23	HQ671736	wnv	HQ671736	2001-XX-X	North Ame	USA	NY	NY	genome	42.1657	-74.9481
24	HQ671737	wnv	HQ671737	2002-XX-X	North Ame	USA	NY	NY	genome	42.1657	-74.9481
25	HQ671738	wnv	HQ671738	2004-XX-X	North Ame	USA	NY	NY	genome	42.1657	-74.9481
26	HQ671739	wnv	HQ671739	2004-XX-X	North Ame	USA	NY	NY	genome	42.1657	-74.9481
27	HQ671740	wnv	HQ671740	2005-XX-X	North Ame	USA	NY	NY	genome	42.1657	-74.9481
28	HQ671741	wnv	HQ671741	2006-XX-X	North Ame	USA	NY	NY	genome	42.1657	-74.9481
29	HQ671743	wnv	HQ671743	2003-XX-X	North Ame	USA	IL	IL	genome	40.3495	-88.9861
30	HQ671744	wnv	HQ671744	2005-XX-X	North Ame	USA	IL	IL	genome	40.3495	-88.9861
31	HQ671745	wnv	HQ671745	2006-XX-X	North Ame	USA	IL	IL	genome	40.3495	-88.9861

- To make your phylogenetic analysis as informative as possible, here are the minimum metadata required.
- Ideally location is more refined than country
- Can add whatever additional metadata you may find useful!
  - E.g. drug resistance phenotype, host/vector species, etc.

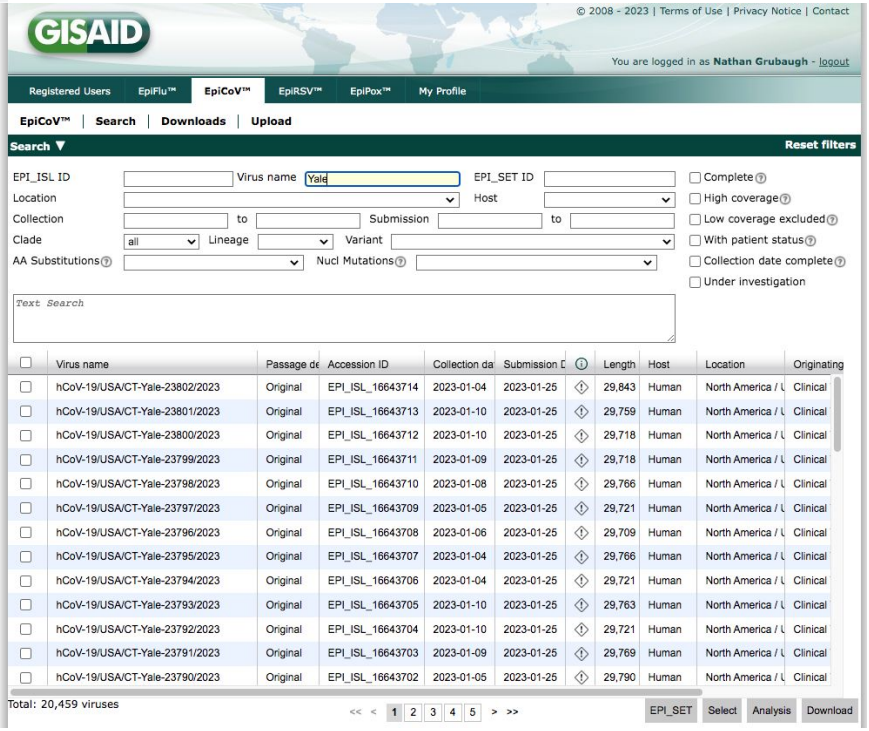
Where to find the metadata?

Depends on your database!

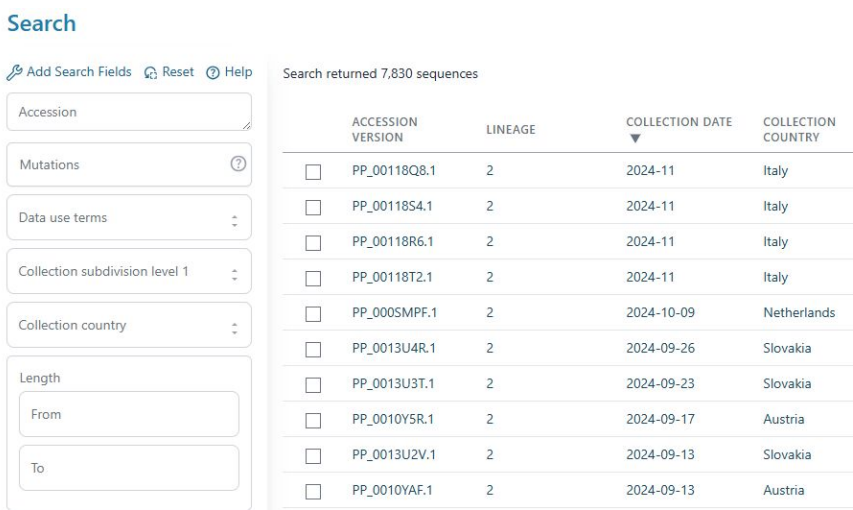
NCBI



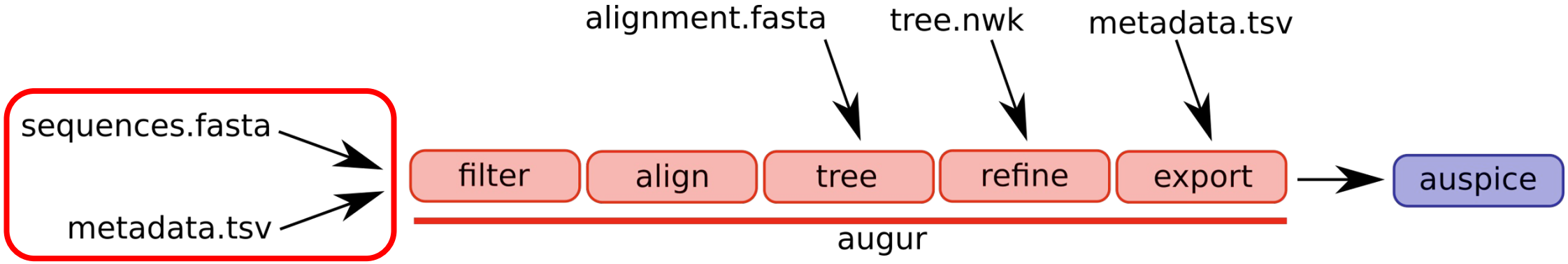
GISAID



Pathoplexus



Preparing to run a Nextstrain build



- **Next-Generation Sequencing Data -> Consensus Sequence Generation**
  - Data types (.fastq, .fasta, .bam)
  - Reference mapping
  - Variant calling
- **Identifying relevant data sources/repositories for contextual data**
  - Pathoplexus
  - NCBI GenBank (SRA, Virus, Microbe)
  - GISAID
  - PlsmoDB
  - BacWGSTdb

Questions?

Reach out via email: jfauver@unmc.edu