Workshop: Online tools for pathogen genome analyses

November 14th, 2023

Sira Sriswasdi, PhD

- Research Affairs, Faculty of Medicine, Chulalongkorn University
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

With Ananporn Supataragul & Apaporn Rodpan

Today's team

Sira Sriswasdi, PhD

Center of Excellence in Computational Molecular Biology
Center for AI in Medicine
Research Affairs

Ananporn Supataragul

Apaporn Rodpan

Plan for today

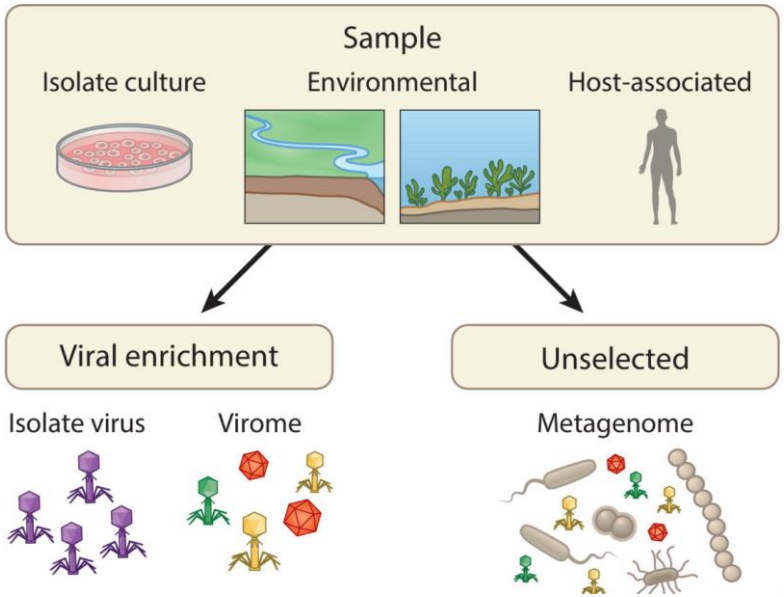
Overview

- Key considerations surrounding pathogen genome analyses
- Some technical points (bioinformatics)
- Introducing online tools

Showcases & Hands-on

- **Nextclade** suite for viral genome analysis
- **Galaxy** bioinformatics platform

How complex is your DNA/RNA sample?



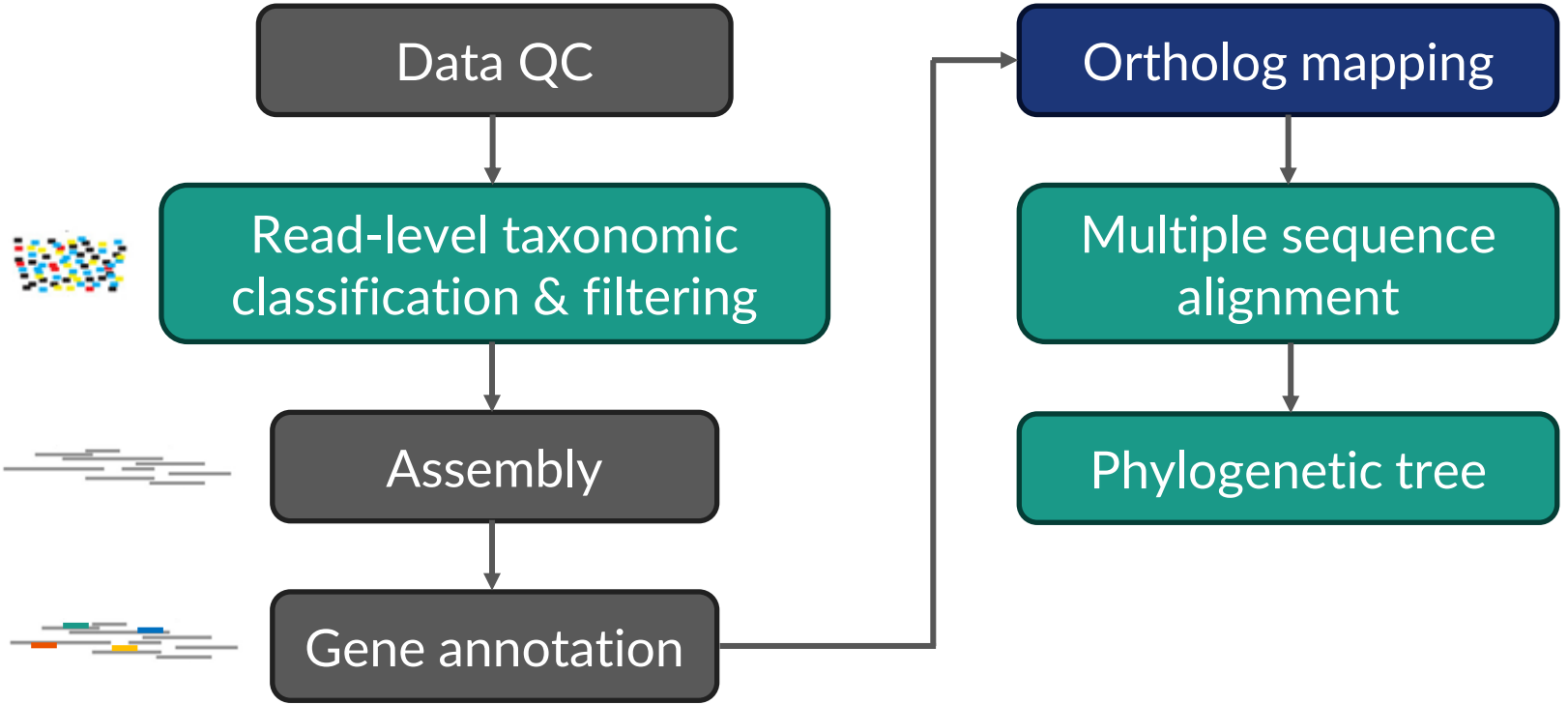
https://jgi.doe.gov/promoting-the-power-of-viral-metagenomics/

- Culture = almost single genome
  - Remove only poor-quality reads
- Host contamination
  - Mark and remove host sequences
- Metagenomics = many organisms
  - Identify appropriate database!

Investigative questions

- **Which strains / variants are present?**
  - Strain / variant calling by aligning to curated DB
  - Taxonomy classification by alignment or *k*-mer matching
- **What are their genomes?**
  - Whole genome reconstruction
  - Novel pathogen identification
- **How are they evolutionally related?**
  - Phylogenetics
  - Visualization

A rough workflow



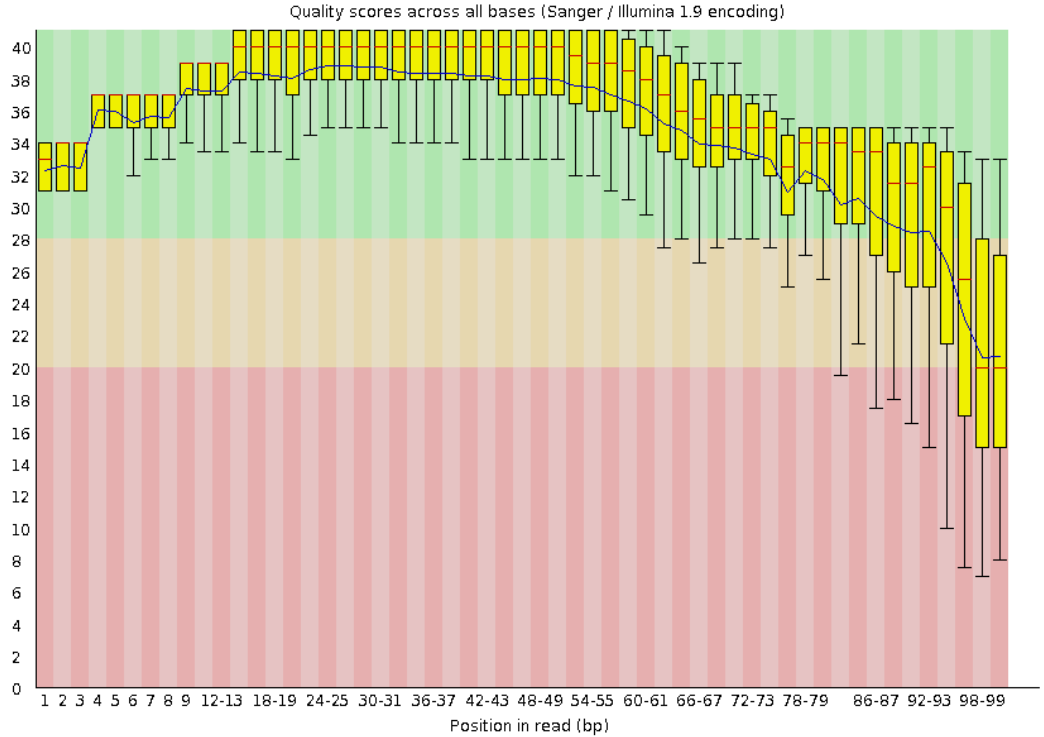
Quality check with FastQC

Basic Statistics

Measure	Value
Filename	small_rna.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	100
%GC	45

- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Per base sequence quality

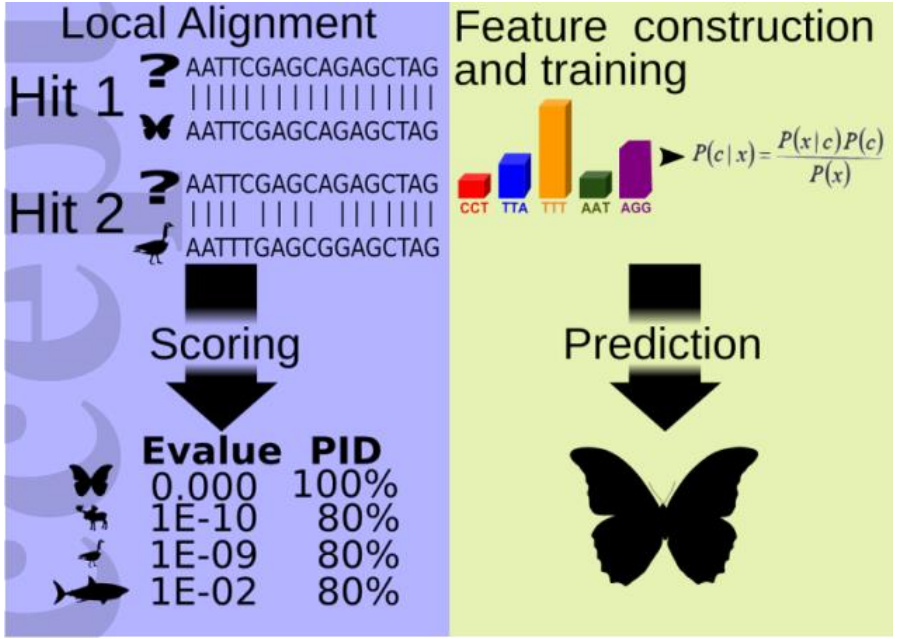


Quality check with FastQC

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
TGAGGTAGTAGATTGTATAGTTAGATCGGAAGAGCACACGTCTGAACTCC	10865	4.346	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TAGCTTATCAGACTGATGTTGACAGATCGGAAGAGCACACGTCTGAACTC	10845	4.338	Illumina Multiplexing PCR Primer 2.01 (100% over 27bp)
TCTTTGGTTATCTAGCTGTATGAGATCGGAAGAGCACACGTCTGAACTCC	7062	2.8247999999999998	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TCTTTGGTTATCTAGCTGTATGAAGATCGGAAGAGCACACGTCTGAACTC	4056	1.6223999999999998	Illumina Multiplexing PCR Primer 2.01 (100% over 27bp)
TGAGGTAGTAGTTTGTGCTGTTAGATCGGAAGAGCACACGTCTGAACTCC	3737	1.4948	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TGAGGTAGTAGTTTGTACAGTTAGATCGGAAGAGCACACGTCTGAACTCC	3549	1.4196	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TGAGGTAGTAGGTTGTATGGTTAGATCGGAAGAGCACACGTCTGAACTCC	2931	1.1724	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
AACCCGTAGATCCGATCTTGTAGATCGGAAGAGCACACGTCTGAACTCCA	1910	0.764	Illumina Multiplexing PCR Primer 2.01 (100% over 29bp)
CGCGACCTCAGATCAGACGTAGATCGGAAGAGCACACGTCTGAACTCCAG	1749	0.6996	Illumina Multiplexing PCR Primer 2.01 (100% over 30bp)
TGAGGTAGTAGGTTGTATAGTTAGATCGGAAGAGCACACGTCTGAACTCC	1647	0.6588	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TCTTTGGTTATCTAGCTGTATAGATCGGAAGAGCACACGTCTGAACTCCA	1622	0.6487999999999999	Illumina Multiplexing PCR Primer 2.01 (100% over 29bp)
TAGCTTATCAGACTGATGTTGATAGATCGGAAGAGCACACGTCTGAACTC	1328	0.5312	Illumina Multiplexing PCR Primer 2.01 (100% over 27bp)
TTCAAGTAATCCAGGATAGGCTAGATCGGAAGAGCACACGTCTGAACTCC	1248	0.4992	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
AGCAGCATTGTACAGGGCTATGAAGATCGGAAGAGCACACGTCTGAACTC	1248	0.4992	Illumina Multiplexing PCR Primer 2.01 (100% over 27bp)

Taxonomic classification



- Sequence alignment against a curated database
- Prediction based on *k*-mer profiles (fast but less accurate)

Hleap, J.S. *et al*. Molecular Ecology Resource, 21(7):2190-2203 (2021)

Phylogenetics

Multiple sequence alignment

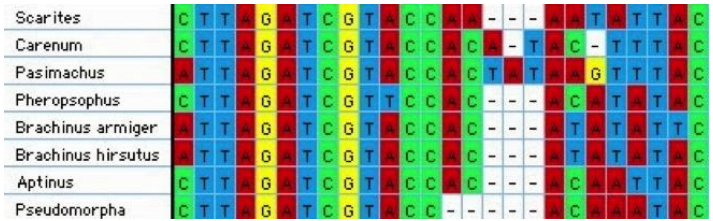
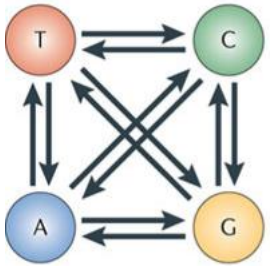


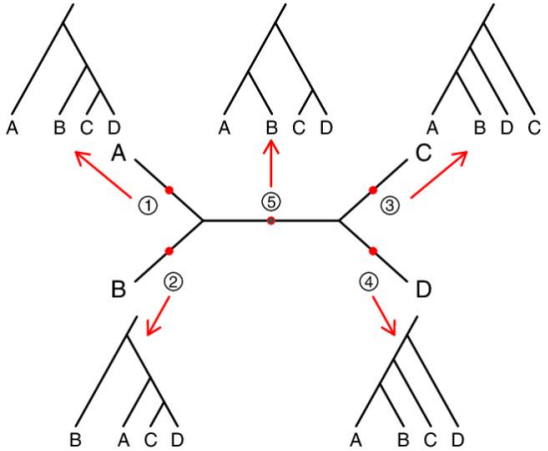
Image from www.mcqbiology.com

Evolutionary model



Yang & Rannala. Nat Rev Genetics (2012)

Tree building algorithm

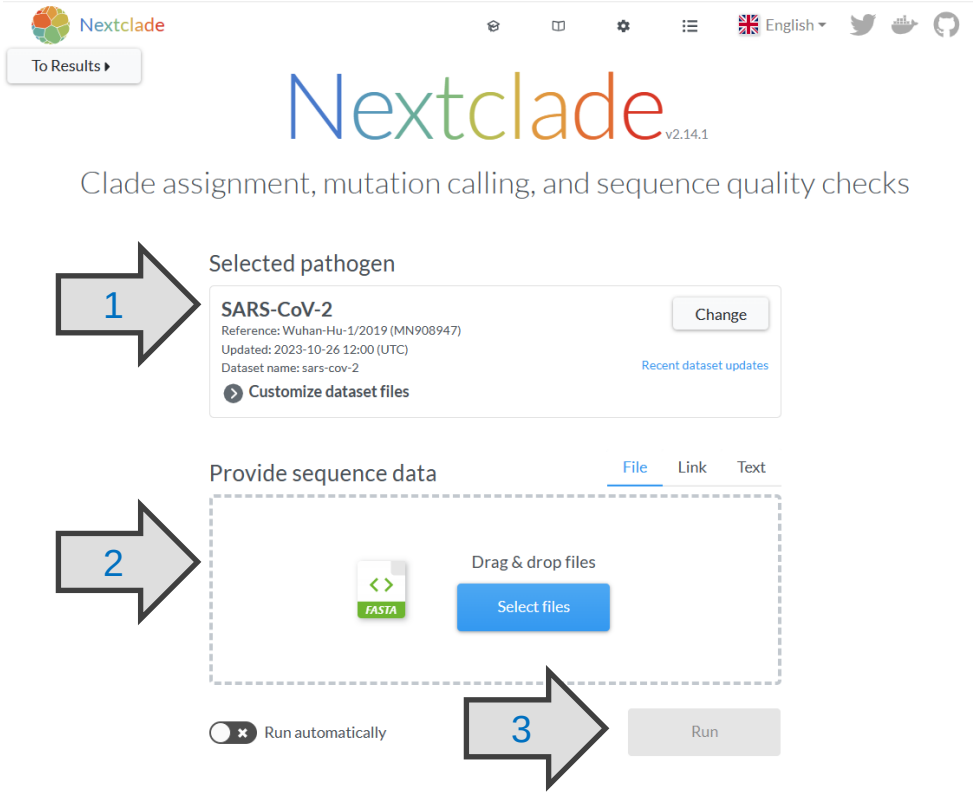


Tian, Y. and Kubatko, L.S. BMC Evol Biol 17(1) (2017)

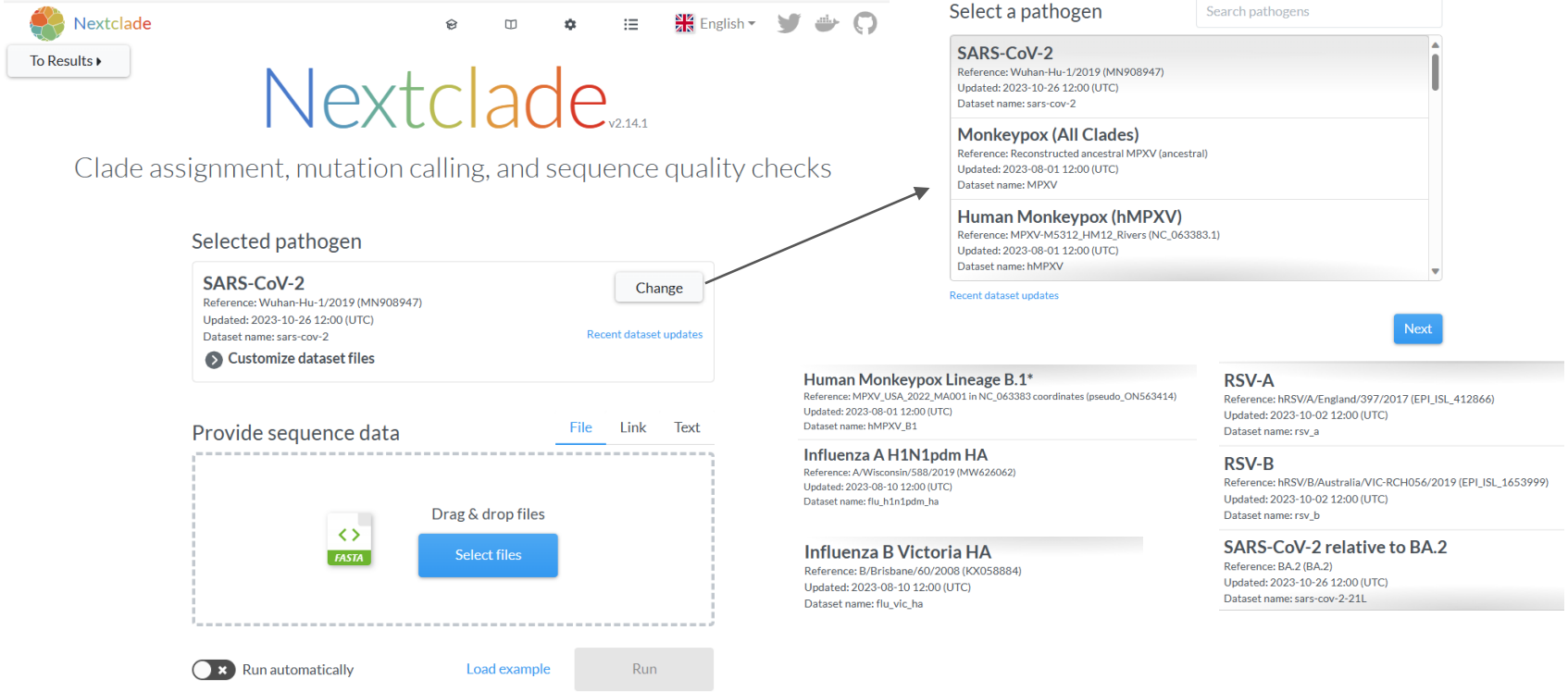
- Estimating the most likely evolutionary relationship given a multiple sequence alignment and a sequence substitution model

Nextclade

Nextclade


Nextclade
English
To Results
Nextclade v2.14.1
Clade assignment, mutation calling, and sequence quality checks
1
Selected pathogen
SARS-CoV-2
Reference: Wuhan-Hu-1/2019 (MN908947)
Updated: 2023-10-26 12:00 (UTC)
Dataset name: sars-cov-2
Change
Recent dataset updates
Customize dataset files
2
Provide sequence data
File
Link
Text
Drag & drop files
FASTA
Select files
Run automatically
3
Run

Supported pathogens

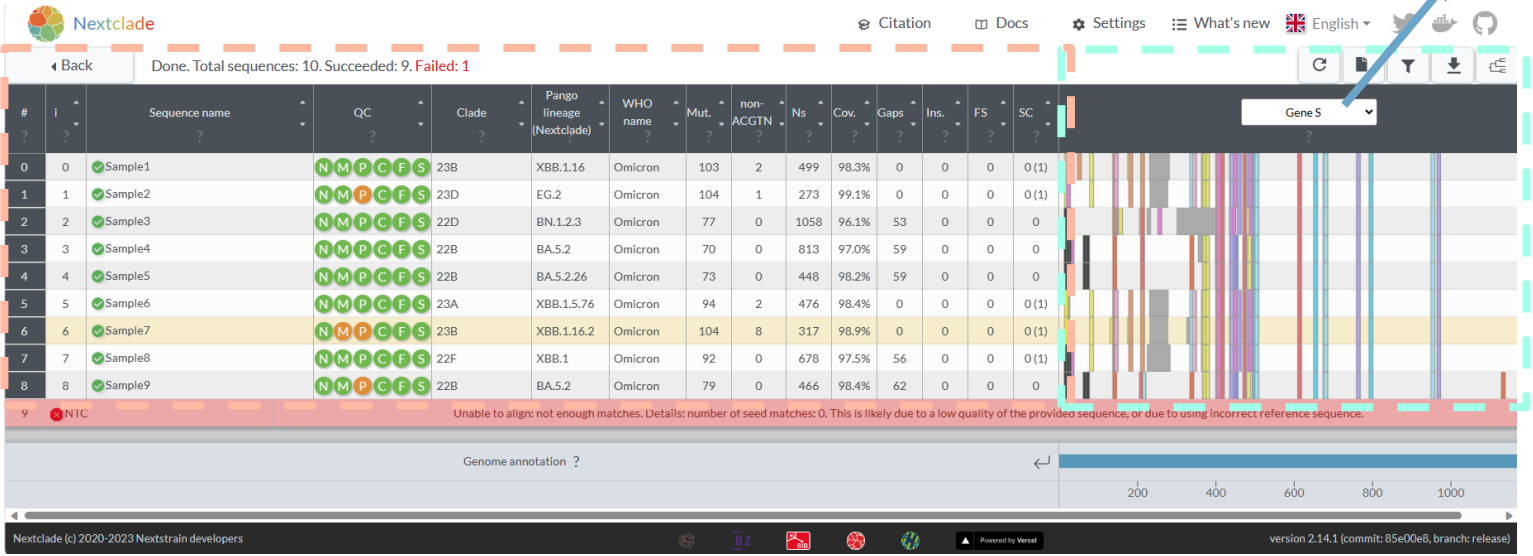


Alignment overview

Sequence classification and quality

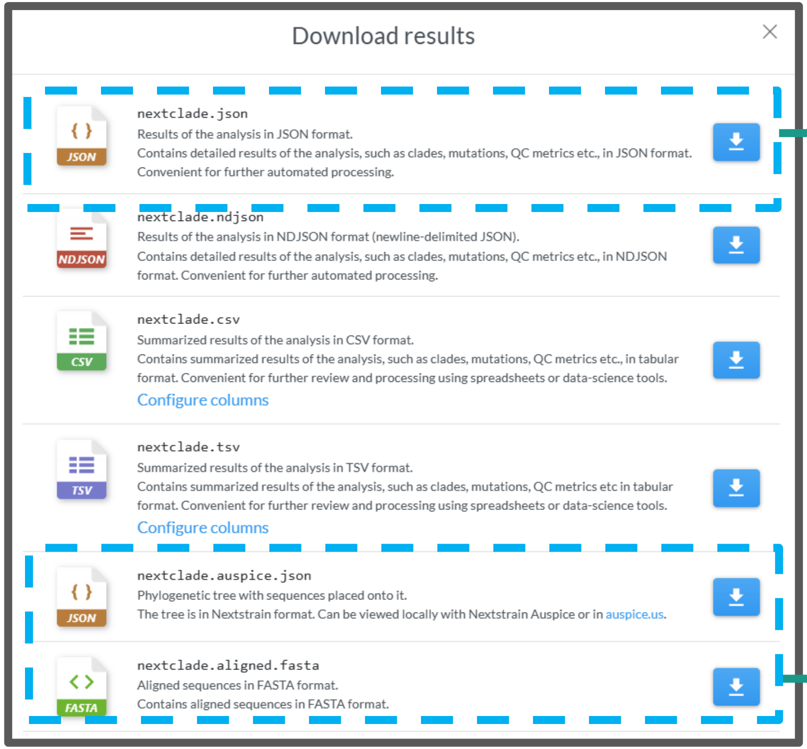
Position of detected mutation against the reference genome

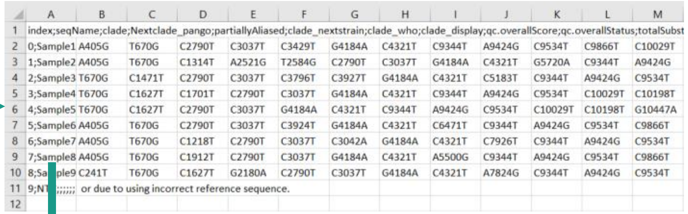
Switch between genes and whole genome

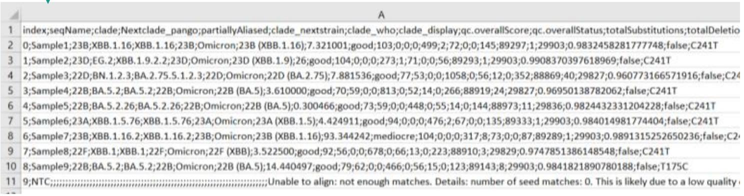


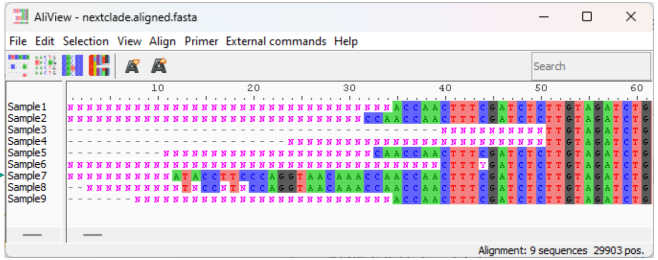
#	i	Sequence name	QC	Clade	Pango lineage (Nextclade)	WHO name	Mut.	non-ACGTN	Ns	Cov.	Gaps	Ins.	FS	SC
0	0	Sample1	N M P C F S	23B	XBB.1.16	Omicron	103	2	499	98.3%	0	0	0	0 (1)
1	1	Sample2	N M P C F S	23D	EG.2	Omicron	104	1	273	99.1%	0	0	0	0 (1)
2	2	Sample3	N M P C F S	22D	BN.1.2.3	Omicron	77	0	1058	96.1%	53	0	0	0
3	3	Sample4	N M P C F S	22B	BA.5.2	Omicron	70	0	813	97.0%	59	0	0	0
4	4	Sample5	N M P C F S	22B	BA.5.2.26	Omicron	73	0	448	98.2%	59	0	0	0
5	5	Sample6	N M P C F S	23A	XBB.1.5.76	Omicron	94	2	476	98.4%	0	0	0	0 (1)
6	6	Sample7	N M P C F S	23B	XBB.1.16.2	Omicron	104	8	317	98.9%	0	0	0	0 (1)
7	7	Sample8	N M P C F S	22F	XBB.1	Omicron	92	0	678	97.5%	56	0	0	0 (1)
8	8	Sample9	N M P C F S	22B	BA.5.2	Omicron	79	0	466	98.4%	62	0	0	0
9		NTC												

Results download

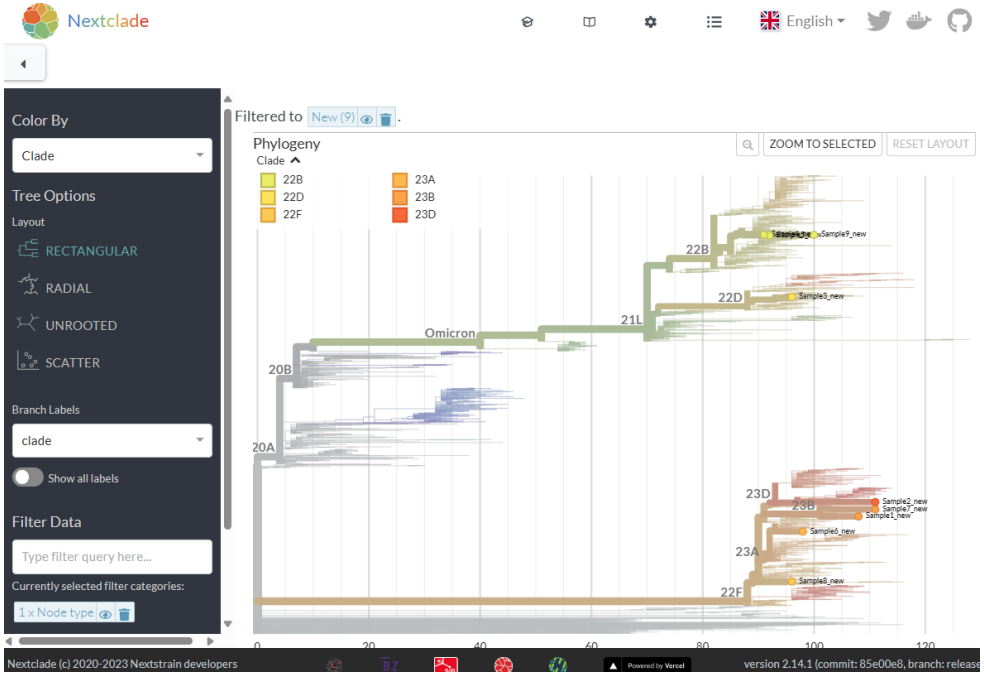


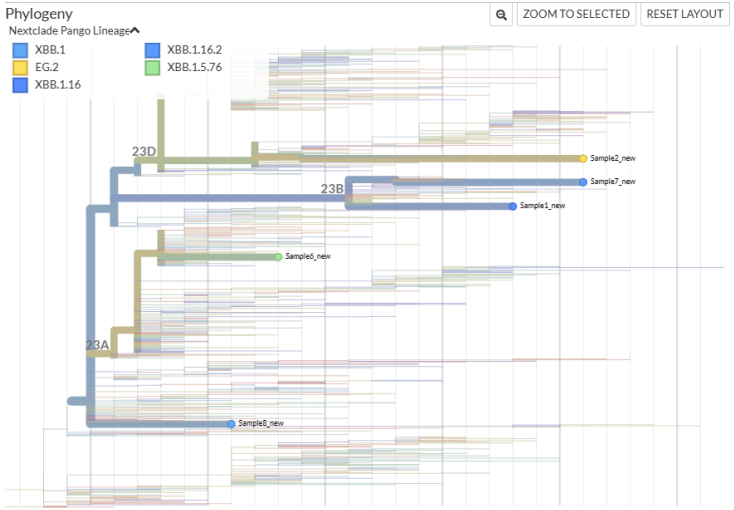


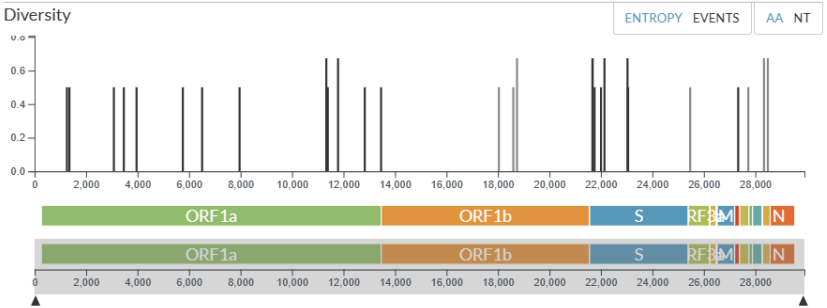




Phylogenetic tree



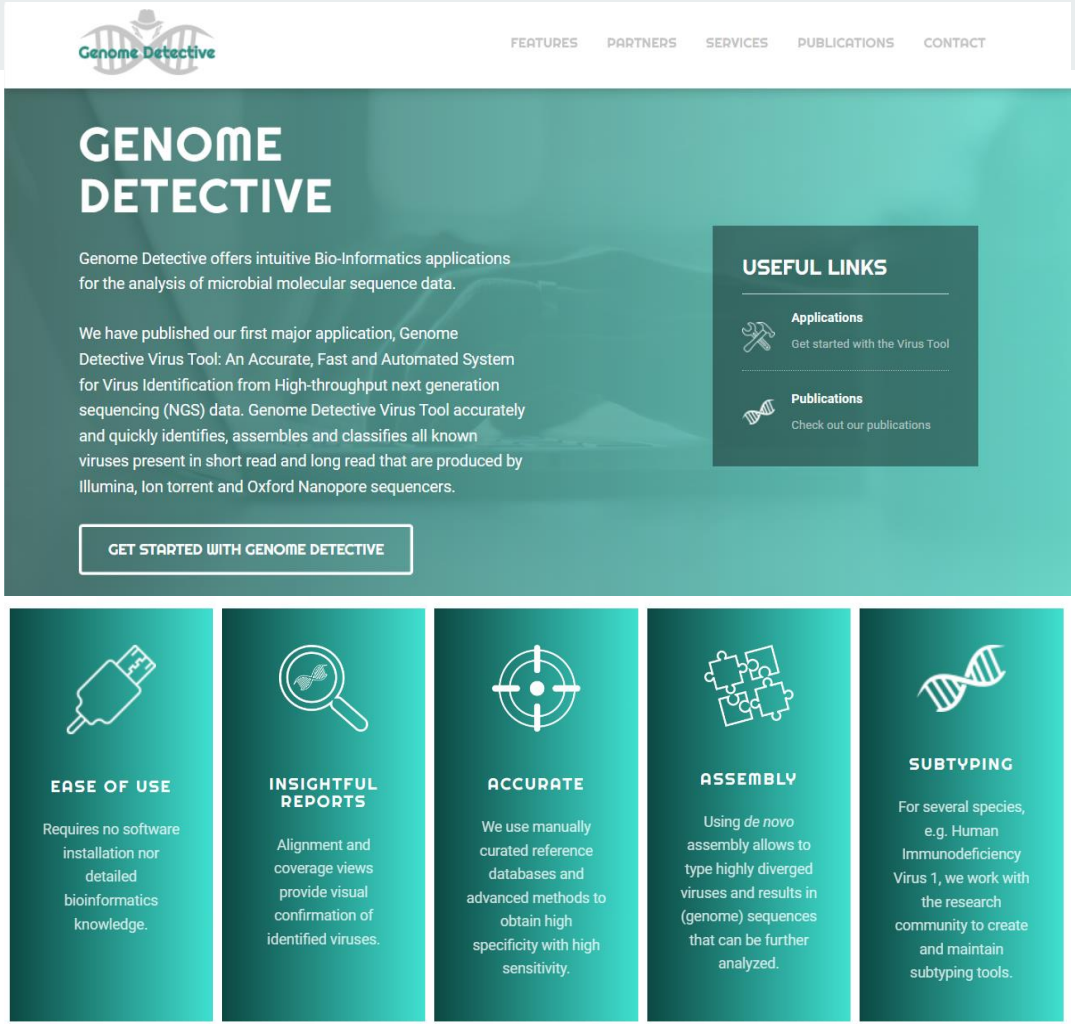




Genome Detective

Genome Detective

- Quality control
- Taxonomy assignment
- Genome assembly
- Subtyping
- Alignment statistics



Data quality check

GENOME DETECTIVE PLATFORM Version 2.13.1

ANALYSIS RESULTS PDF

ANALYSIS OF PEH230633

Technology	Paired-end short reads
Input Files	PEH230633_S5_L001_R1_001.fastq.gz (199.61 MB), PEH230633_S5_L001_R2_001.fastq.gz (207.85 MB)
Protocol	Default Virus Analysis (metagenome)
Submitted On	2023-10-01 15:52:34 UTC
Duration	0h 42m 40s
Release Version	panviral2.64

STATISTICS

Original Read Length	35 - 151
Trimmed Read Length	50 - 136

	# Reads	% of Reads
Input file	4909962	100.0%
After QC	3943492	80.3%
After filtering	88924	1.8%
Mapped back to contigs	68672	1.4%

QUALITY CONTROL (QC) REPORTS

Before Preprocessing	QC report of reads 1	QC report of reads 2
After Preprocessing	QC report of reads 1	QC report of reads 2

FastQC Report

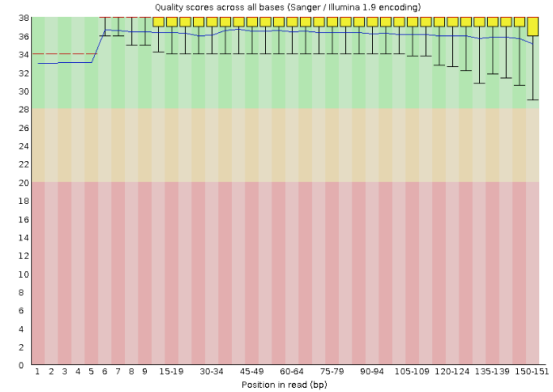
Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Basic Statistics

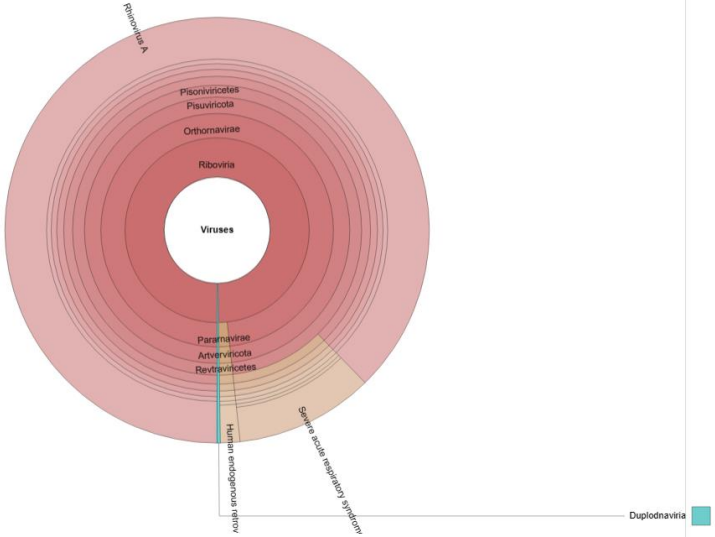
Measure	Value
Filename	PE1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2454981
Sequences flagged as poor quality	0
Sequence length	35-151
%GC	41

Per base sequence quality



Produced by FastQC (version 0.11.5)

Taxonomy chart



Taxonomy chart | Taxonomy tree

Include Discovery ☐

Scaling read count (linear)

ASSIGNMENTS

Strains with at least 50% NT similarity (assuming a 45% similarity for non-covered parts)

Assignment	# Reads	Depth of Coverage	NT Identity	AA Identity	Genome Coverage	Genome Coverage image
Rhinovirus A (subtype: HRV-A68)	57659	976.4	69.2%	72.2%	99.4%	
Severe acute respiratory syndrome-related coronavirus (subtype: SARS-CoV-2, International A_B Diversity)	6940	64.6	99.7%	98.5%	32.1%	
Human endogenous retrovirus K113	977	11.5	99.9%	99.6%	94.8%	
Dubowvirus SAP26	161	3.7	99.3%	98.3%	11.5%	

DISCOVERIES

Strains with low NT identity and/or low genome coverage

Similar to	# Reads	Depth of Coverage	NT Identity	AA Identity	Genome Coverage	Genome Coverage image
Baboon endogenous virus strain M7	156	8.8	58.6%	54.2%	21.8%	
Peeveelvirus PV83	130	5.4	90.2%	87.0%	5.6%	
Triavirus P240	115	3.2	99.6%	99.0%	8.7%	
Human coronavirus 229E	105	52.2	86.1%	79.2%	0.6%	

Assembly and annotation

GENOME DETECTIVE PLATFORM

Version 2.13.1

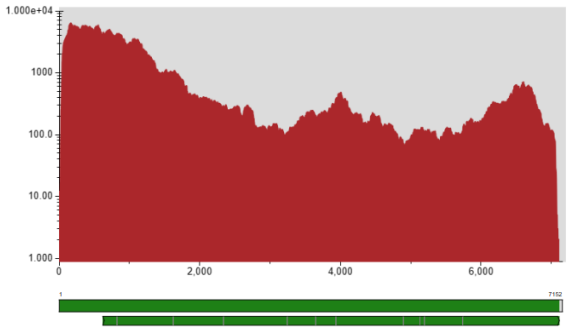
NGS DETAILS (PEH230633): RHINOVIRUS A

ASSEMBLY

Coverage Length	7111 (1 contig(s))
Depth Of Coverage	976.4
Number Of Reads	57659
Reads Per Million	14621.31 rpm (after QC)
Ambiguities	0
Assembly Method	de novo + reference guided assembly
Consensus Caller	GATK

COVERAGE MAP

Map



ASSIGNMENT

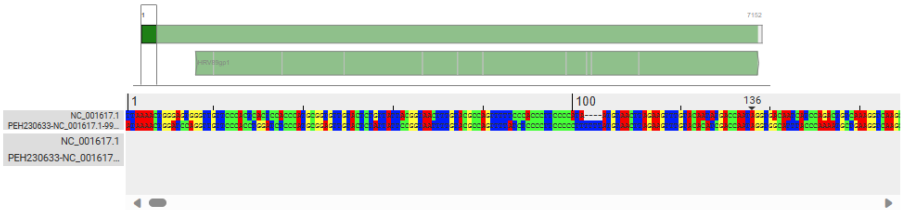
Type	Rhinovirus A (Taxonomy ID: 147711)
Subtype	HRV-A68
Reference Genome	NC_001617.1
NT Identity (%)	69.193
AA Identity (%)	72.1735
Number Of Stop Codons	0
Number Of CDS	1

ALIGNMENT

Alignment Score	5322.0 (NT) + 11678.0 (AA) = 17000.0
Concordance (%)	58.2431
Alignment Method	Global, seeded, nucleotide + amino acids (AGA)
NT Alignment	Download alignment (FASTA)
Region Bam File	Download BAM file (.bam)
CDS Alignment	Download CDS alignment (FASTA)
Consensus Sequence	Download consensus sequence (FASTA)
Consensus Contigs	Download consensus contigs (FASTA)

GENOME REGION

Sequence starts at position 1 and ends at position 7111 relative to NC_001617.1 reference sequence.



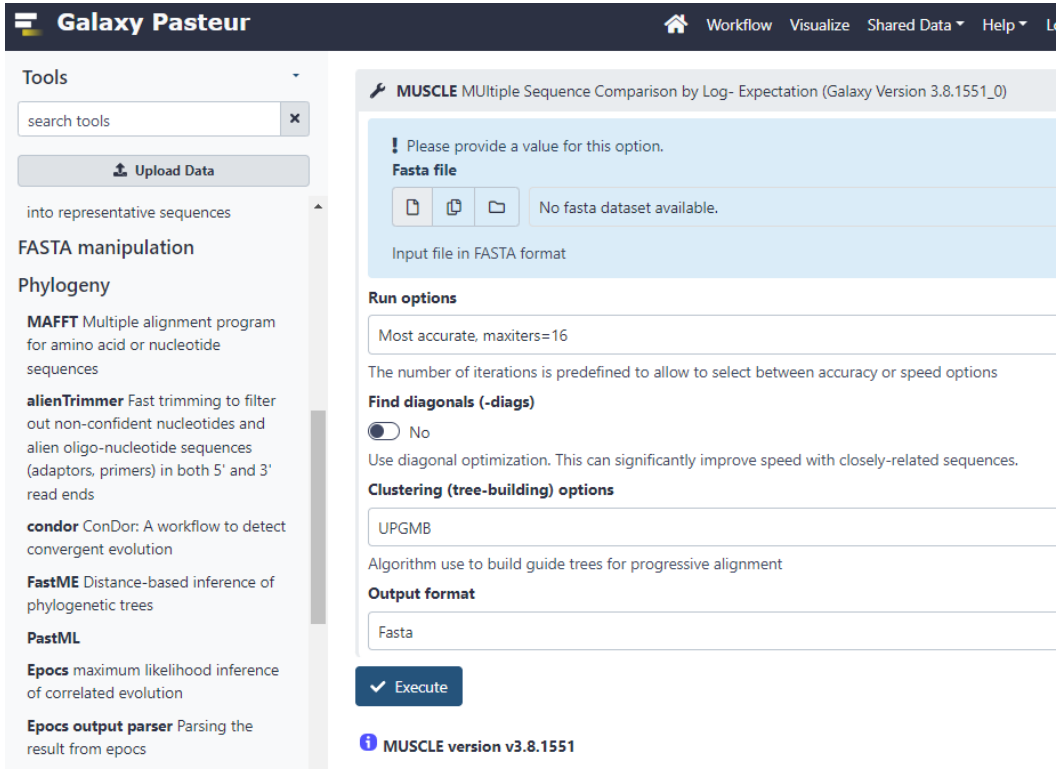
ALIGNMENT DETAILED STATISTICS

SHOW MUTATIONS

	Begin	End	Coverage	Score	Concordance	Matches	Identities	I/D/M/F*	Stop Codons
NT	1	7111	99.4%	5322	37.8%	7082 (99.0%)	4930 (68.9%)	43/29	
CDS									
HRV89gp1	1	2165	100%	11678	77.2%	2156 (99.1%)	1564 (71.9%)	11/9/0/0	0
Proteins									
genome polypr...	1	2165	100%	11678	77.2%	2156 (99.1%)	1564 (71.9%)	11/9/0/0	0
capsid protein ...	1	69	100%	436	95.8%	69 (100%)	66 (95.7%)	0/0/0/0	0

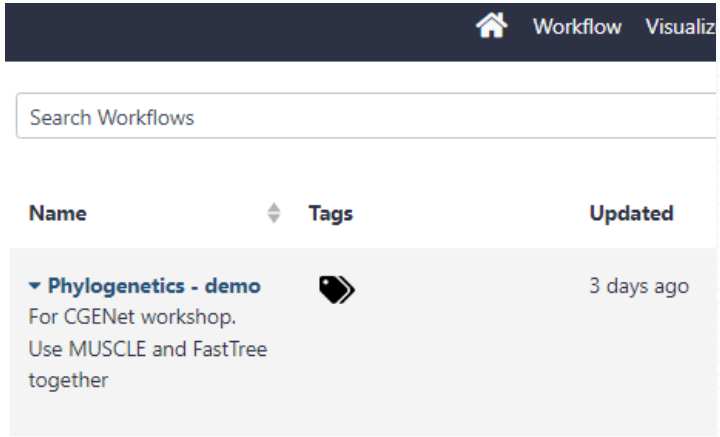
Galaxy

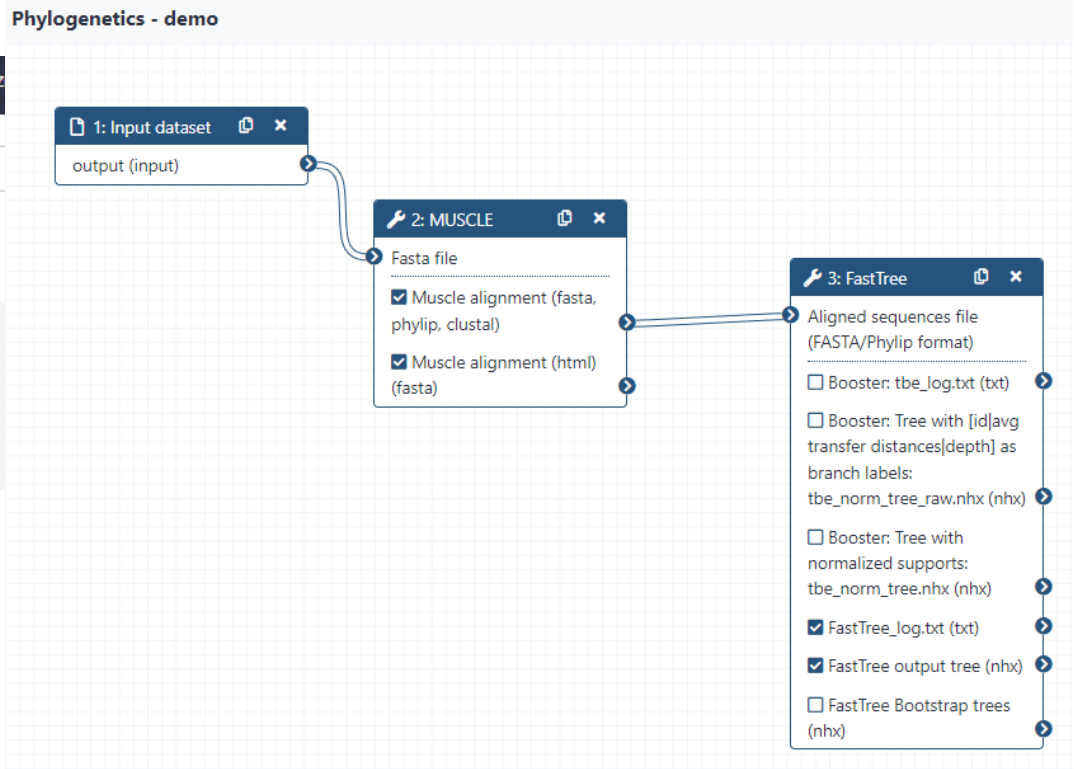
Galaxy



- A GUI platform with many bioinformatics tools built-in
- Upload data file and set the analysis parameters
  - Require bioinformatics understanding but **no coding skill**

Customizable workflow





- Multiple bioinformatics tools can be chained together into a workflow

Bioinformatics parameter setting

Step 2: MUSCLE

Fasta file
Output dataset 'output' from step 1

Run options
Most accurate, maxiters=16

Find diagonals (-diags)
False

Clustering (tree-building) options
UPGMB

Output format
Fasta

Step 3: FastTree

Aligned sequences file (FASTA/Phylip format)
Output dataset 'out_align' from step 2

Nucleotide or protein alignment
Auto detect

Evolutionary model if type is DNA
GTR

Evolutionary model if type is PROTEIN
LG

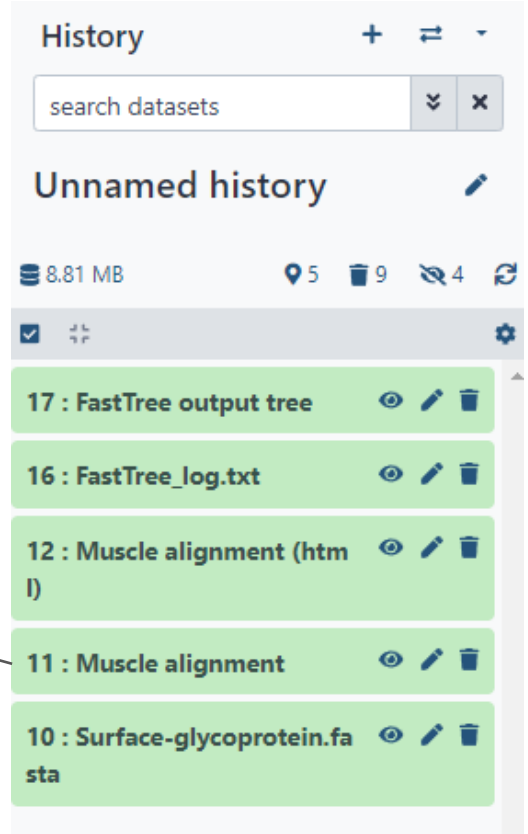
Use Gamma distribution
Yes

Bootstrap branch support
Bootstrap

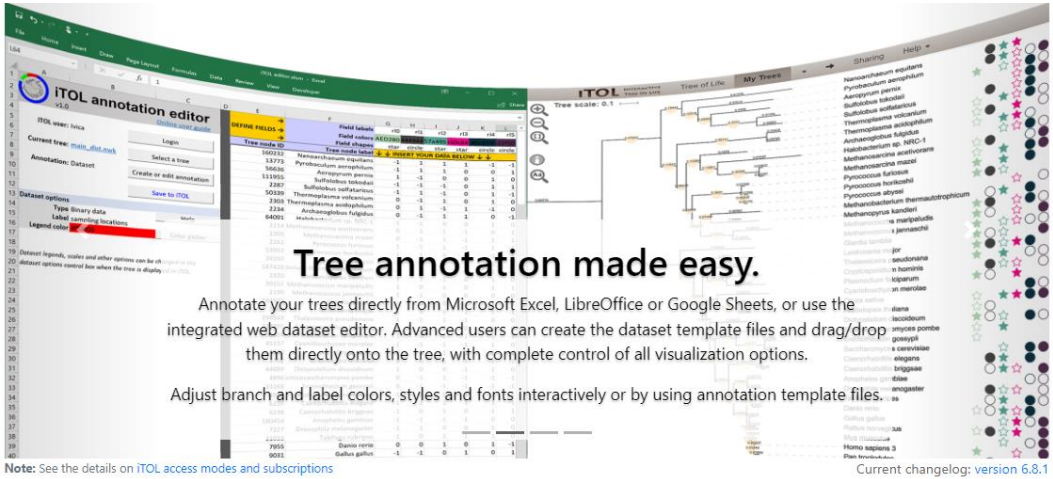
Number of replicates
100

Analysis history

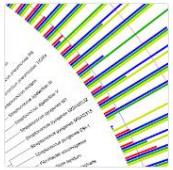
```
>lcl|NC_055229.1_cds_YP_010085331.1_1 [locus_tag=KM541_gp045] [db_xref=GeneID:65100403]
[location=42566..48658] [gbkey=CDS]
ATGA------------AGCAACTCGGGGCGCTTGCGCTAGCGCTCGCGCTGGTGTTCGCG
TTCGTGCCCGTACTGCACGCGAACGCCAGCAAGTGCTACCGCAAGTCAGGATTCTACCAC
CAT---------------GGTGAAAAACAAGGCGGCGGCAGCGGCGGCAGCGGCGGCAGC
GGCAACGAGCCTCGA----------------------------AACTCGTACGCTGACTT
CATAGACAAGAAACAAGCTTTTCGCAAGATACTC------------------GTCGCTTC
TATCAACTGGGACGATGTCAGGAACACGCTCAAGGAGGAGTTTATCGGAGAGTGCCGCCA
GAGAA-----------------ACGGGTCATCATCAGGATATCTGTACGACTA-------
--------CTCCGCAGCCTTAAACCTGACGCTGCGTGCCGCAGAACTGCGCCGCGGTAAC
CGCTCGCAAGCCGAGAACGCTACGATCATACTCGAATTCGATACGGCTAACGTCACCGTA
GG----------------------------------------------------------
----------------------------------------------------TTTTGT
CTGTAATGCTTCGGTTGCGGAAGTCTCGTAC-CGCATCGTCGGCGATACCGC--------
-----------------TTCTTCGTCGGCGATGCGTACGAACTCTTCTGAAATGT-----
----CTGTTGAGGTGACCTTCTCCGGGATATC------------GGTAGCGC--------
-----------------AACCTTACATGAGATCCGCATCCTGCACGGATGACTGTTTCG
AAATAGCCACCGAAGCCATCTATCCTACAGGCGAACCGTTTACCTCAGCAGGTACAGACA
CTCCCAACGCCACGGCGGAAGCTAC---TACAGCAGAA--------------------GC
TACTACTCCAGCACCTACCACTCCTGTAC----------CTACCACTCC-----------
----CGTACCCTACACGCCTGTT-CCAGAAGCGCGTCGCGGAGCGGTGAT-CATCTCTGA
```



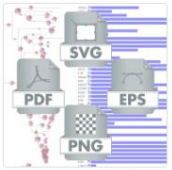
Phylogenetic tree viewer



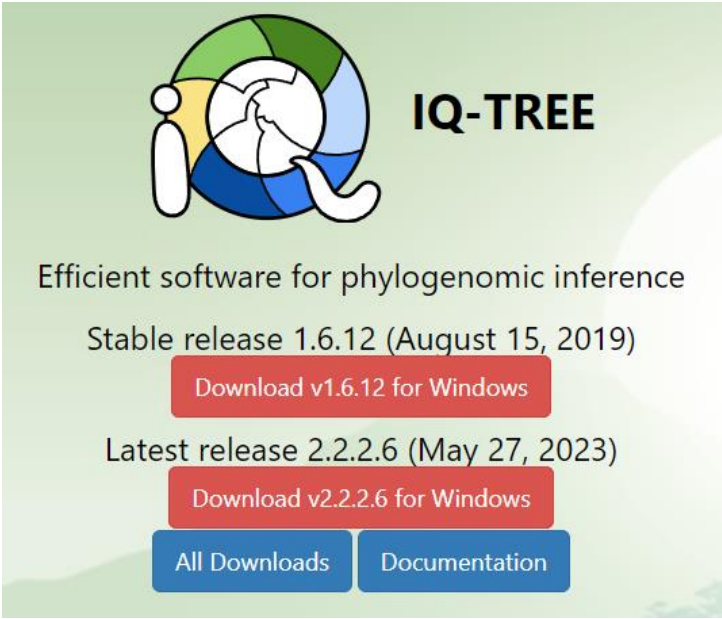
Manage



Annotate



Export



Let’s try analyzing some data

https://github.com/cmb-chula/CGENet-symposium-2023

Thank you for coming today

Contact information

- Sira Sriswasdi: sira.sr@chula.ac.th

Acknowledgements

- Faculty of Medicine and Faculty of Engineering, Chulalongkorn University
- King Chulalongkorn Memorial Hospital, Thai Red Cross Society
- World Health Organization
- The National Research Council of Thailand
- Department of Disease Control, Ministry of Public Health
- All speakers: Mahidol University, the University of Tokyo, Kasetsart University
- CGENet developers

Genomics

Transcriptomics

Proteomics

Metabolomics

Multi-omics

Epigenomics

Single-cell

Nutrigenomics

Micro/Metagenomics

System Biology

Pharmacogenomics

Ai & ML

Infectious Diseases

Agriculture

Others