



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



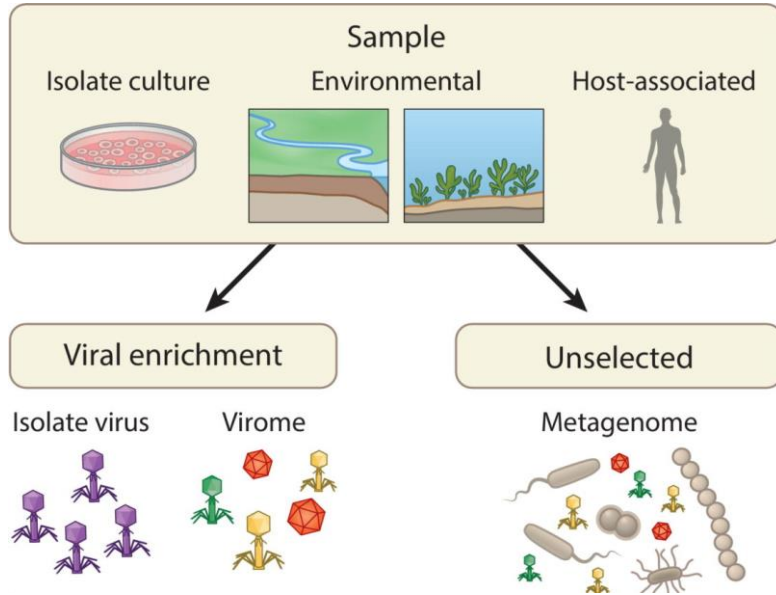
Overview

- ❑ 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?



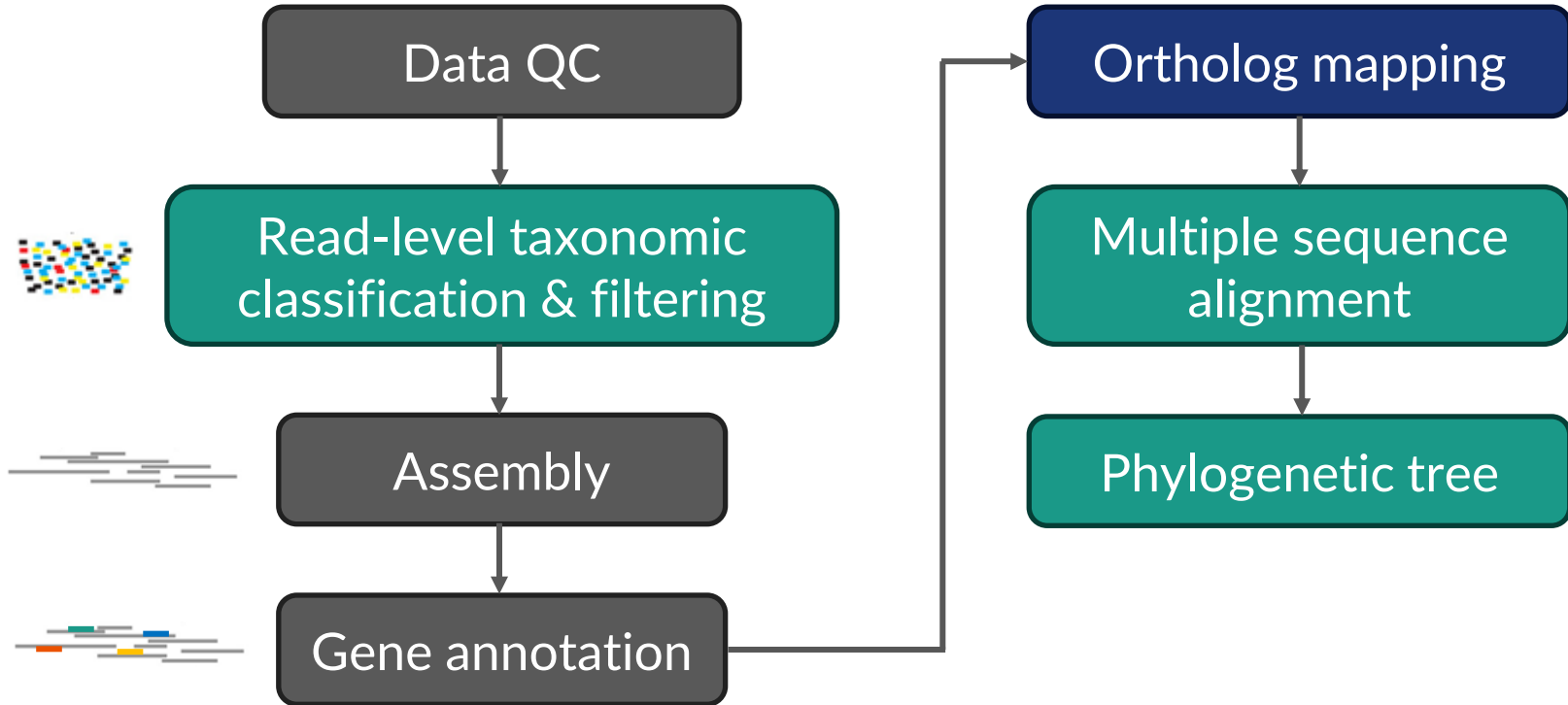
- ❑ Culture = almost single genome
 - ❑ Remove only poor-quality reads
- ❑ Host contamination
 - ❑ Mark and remove host sequences
- ❑ Metagenomics = many organisms
 - ❑ Identify appropriate database!

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



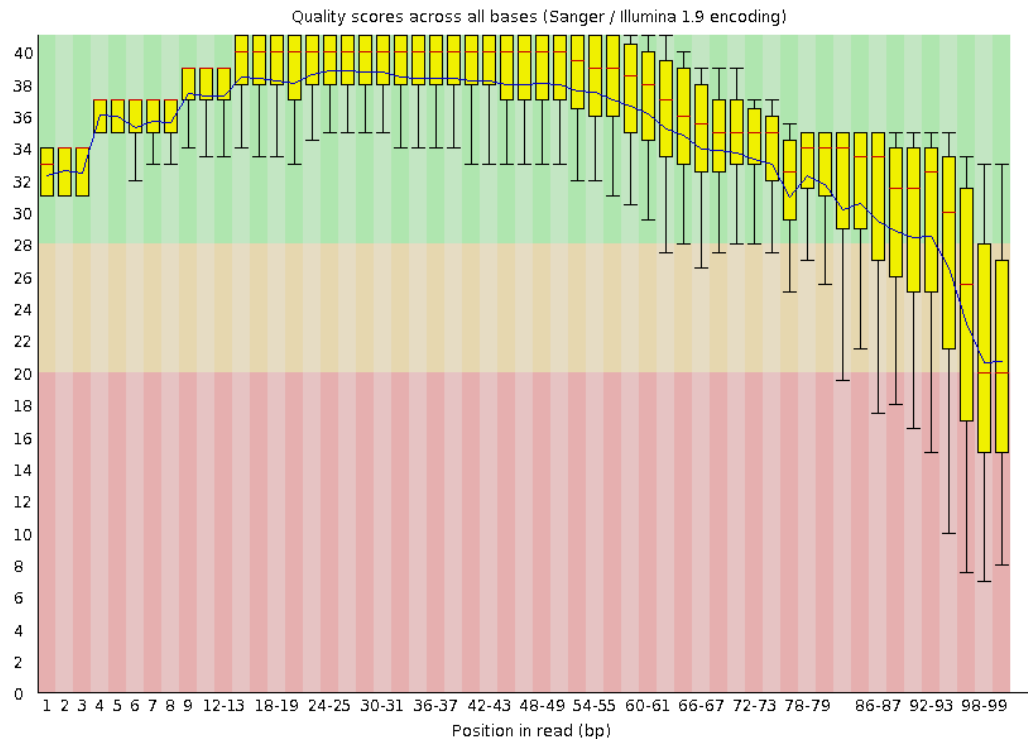
Quality check with FastQC

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

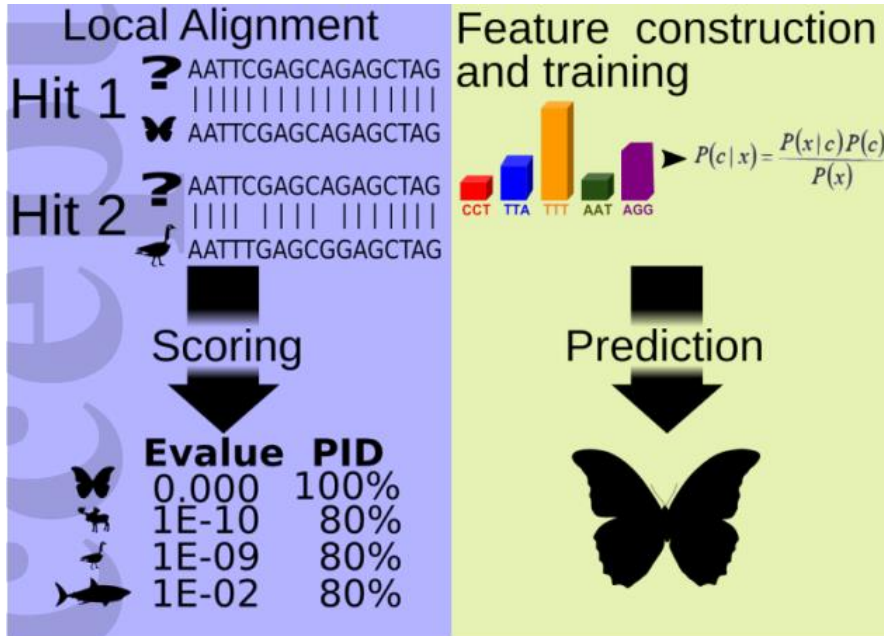


Quality check with FastQC

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	2.824799999999998	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TCTTTGGTTATCTAGCTGTATGAAGATCGGAAGAGCACACGTCTGAACTC	4056	1.622399999999998	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)
TGAGGTAGTAGTTGTATGGTTAGATCGGAAGAGCACACGTCTGAACTCC	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)
TGAGGTAGTAGTTGTATAGTTAGATCGGAAGAGCACACGTCTGAACTCC	1647	0.6588	Illumina Multiplexing PCR Primer 2.01 (100% over 28bp)
TCTTTGGTTATCTAGCTGTATAGATCGGAAGAGCACACGTCTGAACTCCA	1622	0.648799999999999	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)

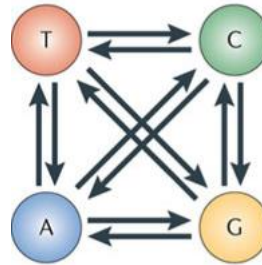
Phylogenetics

Multiple sequence alignment

Scarites	C	T	T	A	G	A	T	C	G	T	A	C	C	A	-	-	-	A	T	T	A	C				
Carenum	C	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	T	A	C	-	T	T	A	C		
Pasimachus	A	T	T	A	G	A	T	C	G	T	A	C	C	A	C	T	A	T	A	G	T	T	T	A	C	
Pheropsophus	C	T	T	A	G	A	T	C	G	T	T	C	C	A	C	-	-	-	A	C	A	T	A	T	A	C
Brachinus armiger	A	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	T	A	T	A	T	T	C
Brachinus hirsutus	A	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	T	A	T	A	T	A	C
Aptinus	C	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	C	A	T	A	T	A	C
Pseudomorpha	C	T	T	A	G	A	T	C	G	T	A	C	C	A	C	-	-	-	A	C	A	T	A	T	A	C

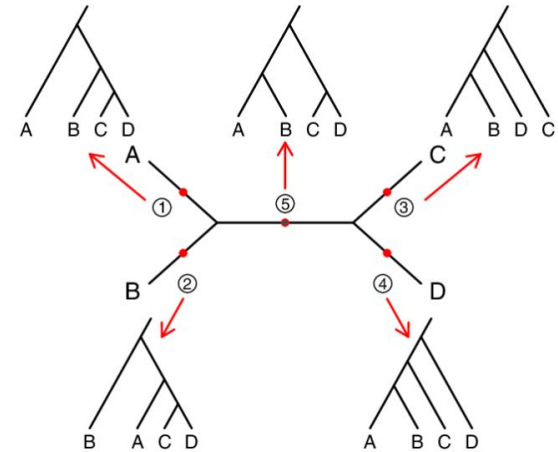
Image from www.mcqbiology.com

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

FileLinkText



FASTA

Drag & drop files

Select files

3

☒ Run automatically

Run

Supported pathogens



Clade assignment, mutation calling, and sequence quality checks

Selected pathogen

SARS-CoV-2

Reference: Wuhan-Hu-1/2019 (MN908947)
Updated: 2023-10-26 12:00 (UTC)
Dataset name: sars-cov-2

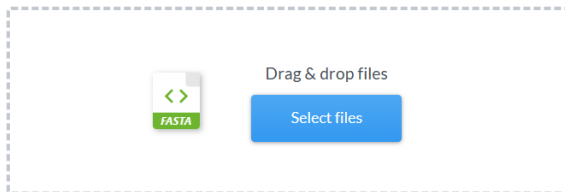
Change

[Recent dataset updates](#)

[Customize dataset files](#)

Provide sequence data

[File](#) [Link](#) [Text](#)



☐ Run automatically

[Load example](#)

[Run](#)

Select a pathogen

SARS-CoV-2

Reference: Wuhan-Hu-1/2019 (MN908947)
Updated: 2023-10-26 12:00 (UTC)
Dataset name: sars-cov-2

Monkeypox (All Clades)

Reference: Reconstructed ancestral MPXV (ancestral)
Updated: 2023-08-01 12:00 (UTC)
Dataset name: MPXV

Human Monkeypox (hMPXV)

Reference: MPXV-M5312_HM12_Rivers (NC_063383.1)
Updated: 2023-08-01 12:00 (UTC)
Dataset name: hMPXV

[Recent dataset updates](#)

[Next](#)

Human Monkeypox Lineage B.1*

Reference: MPXV_USA_2022_MAO01 in NC_063383 coordinates (pseudo_ON563414)
Updated: 2023-08-01 12:00 (UTC)
Dataset name: hMPXV_B1

Influenza A H1N1pdm HA

Reference: A/Wisconsin/588/2019 (MW626062)
Updated: 2023-08-10 12:00 (UTC)
Dataset name: flu_h1n1pdm_ha

Influenza B Victoria HA

Reference: B/Brisbane/60/2008 (KX058884)
Updated: 2023-08-10 12:00 (UTC)
Dataset name: flu_vic_ha

RSV-A

Reference: hRSV/A/England/397/2017 (EPI_ISL_412866)
Updated: 2023-10-02 12:00 (UTC)
Dataset name: rsv_a

RSV-B

Reference: hRSV/B/Australia/VIC-RCH056/2019 (EPI_ISL_1653999)
Updated: 2023-10-02 12:00 (UTC)
Dataset name: rsv_b

SARS-CoV-2 relative to BA.2

Reference: BA.2 (BA.2)
Updated: 2023-10-26 12:00 (UTC)
Dataset name: sars-cov-2-21L

Alignment overview

Sequence classification
and quality

Position of detected
mutation against the
reference genome

Switch between genes
and whole genome

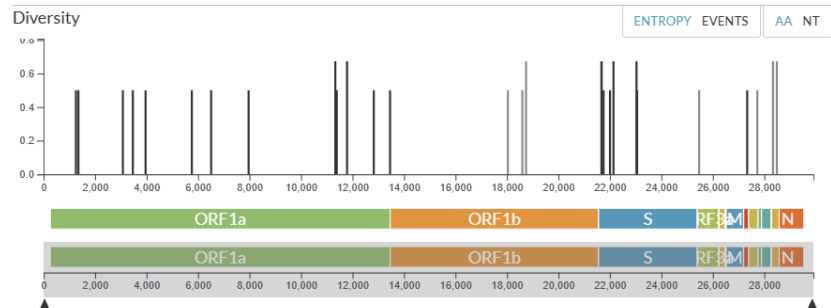
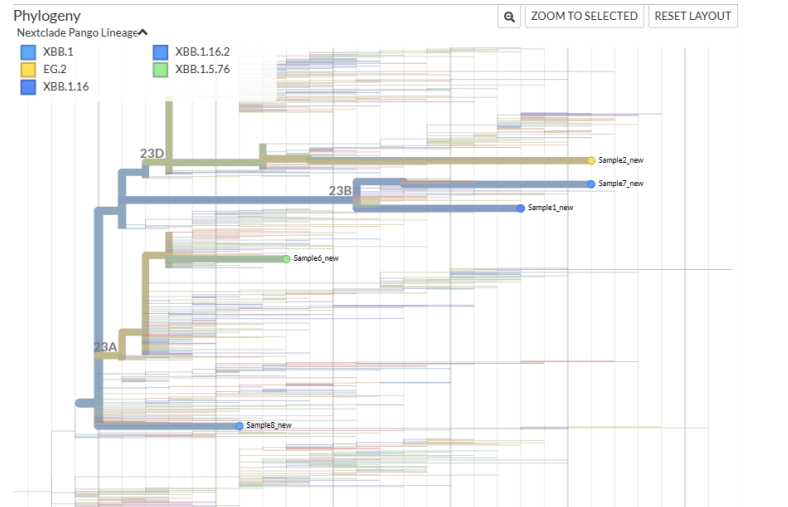
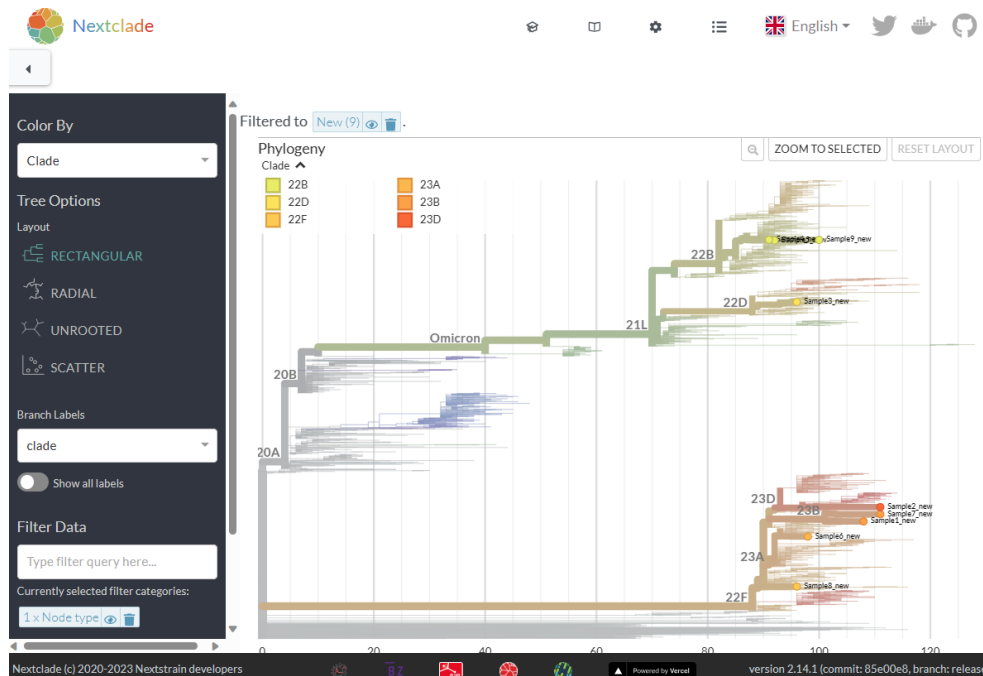


1	index:sa
---	----------

100

The screenshot shows the AllView application window titled "AllView - nextclade.aligned.fasta". The menu bar includes File, Edit, Selection, View, Align, Primer, External commands, and Help. Below the menu is a toolbar with icons for file operations, sequence editing, and alignment. A search bar is located on the right side of the toolbar. The main display area shows a sequence alignment of 9 samples (Sample1 to Sample9) across 60 positions. The alignment is visualized with colored bars representing nucleotide bases (A, C, G, T) and gaps. The alignment is highly conserved, with most positions showing the same base across all samples. The status bar at the bottom right indicates "Alignment: 9 sequences 29903 pos."

Phylogenetic tree






Genome Detective

Genome Detective

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

Genome Detective

FEATURES PARTNERS SERVICES PUBLICATIONS CONTACT


GENOME DETECTIVE


Genome Detective offers intuitive Bio-Informatics applications for the analysis of microbial molecular sequence data.


We have published our first major application, Genome Detective Virus Tool: An Accurate, Fast and Automated System for Virus Identification from High-throughput next generation sequencing (NGS) data. Genome Detective Virus Tool accurately and quickly identifies, assembles and classifies all known viruses present in short read and long read that are produced by Illumina, Ion torrent and Oxford Nanopore sequencers.

GET STARTED WITH GENOME DETECTIVE

USEFUL LINKS


**Applications**
Get started with the Virus Tool

**Publications**
Check out our publications




EASE OF USE

Requires no software installation nor detailed bioinformatics knowledge.




INSIGHTFUL REPORTS

Alignment and coverage views provide visual confirmation of identified viruses.




ACCURATE

We use manually curated reference databases and advanced methods to obtain high specificity with high sensitivity.



ASSEMBLY

Using *de novo* assembly allows to type highly diverged viruses and results in (genome) sequences that can be further analyzed.



SUBTYPING

For several species, e.g. Human Immunodeficiency Virus 1, we work with the research community to create and maintain subtyping tools.

Data quality check

GENOME DETECTIVE PLATFORM

Version 2.13.1

ANALYSIS RESULTS



ANALYSIS OF PEH230633

Technology Paired-end short reads

Input Files PEH230633_SS_L001_R1_001.fastq.gz (199.61 MB),
PEH230633_SS_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 parviral2.64

STATISTICS

Original Read Length 35 - 151

Trimmed Read Length 50 - 136

	# Reads	% of Reads
Input file	4909962	100.0%
After QC 1	3943492	80.3%
After filtering 1	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

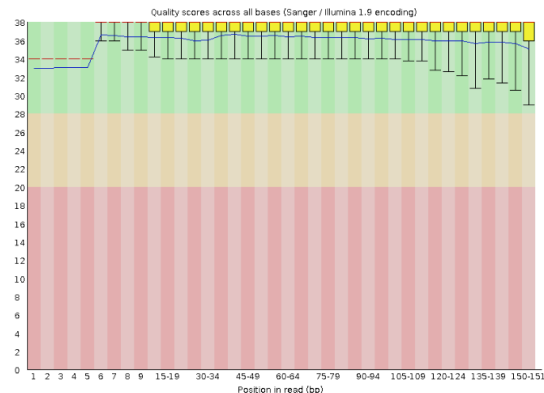
Summary

- ✓ 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

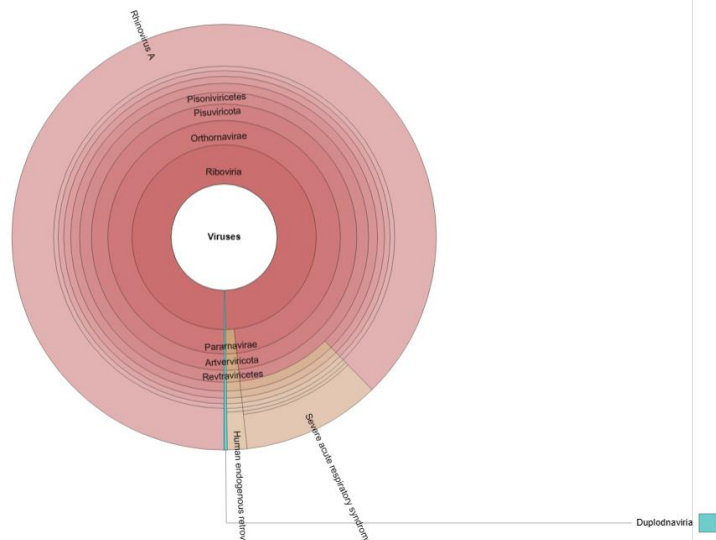
Measure	Value
Filename	PE1.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	2454981
Sequences flagged as poor quality	0
Sequence length	35-151
%GC	41

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%	
Dubovirus 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

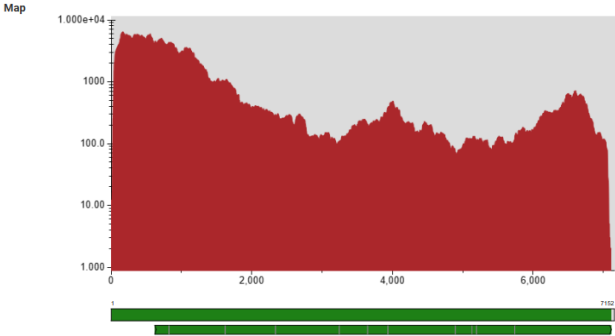
Version 2.13.1

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



ASSIGNMENT

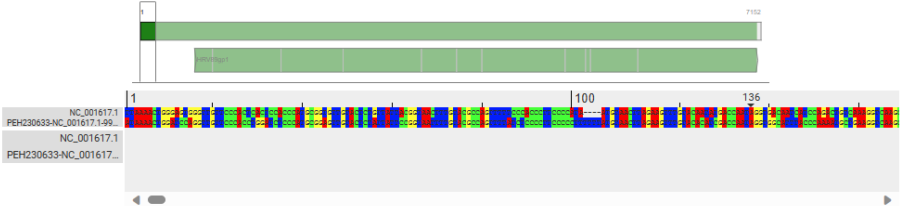
Type	Rhinovirus A (Taxonomy ID: 147711)
Subtype	HRV-A68
Reference Genome	NC_001617.1
NT Identity (%)	69.193
AA Identity (%)	72.1735
Number Of Stop Codons	0
Number Of CDS	1

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									
<input type="checkbox"/> HRV89gp1	1	2165	100%	11678	77.2%	2156 (99.1%)	1564 (71.9%)	11/9/0/0	0
Proteins									
<input type="checkbox"/> genome polypr...	1	2165	100%	11678	77.2%	2156 (99.1%)	1564 (71.9%)	11/9/0/0	0
<input type="checkbox"/> capsid protein ...	1	69	100%	436	95.8%	69 (100%)	66 (95.7%)	0/0/0/0	0



Galaxy

Galaxy

The screenshot displays the Galaxy Pasteur web interface. The top navigation bar includes a home icon, 'Workflow', 'Visualize', 'Shared Data', and 'Help'. On the left, a sidebar lists various tools under categories like 'FASTA manipulation' and 'Phylogeny'. The main area shows the 'MUSCLE' tool configuration page. It includes a 'Fasta file' input section with a message 'No fasta dataset available.' and a 'Run options' section with a dropdown set to 'Most accurate, maxiters=16'. Other sections include 'Find diagonals (-diags)' with a 'No' radio button, 'Clustering (tree-building) options' with a dropdown set to 'UPGMB', and 'Output format' set to 'Fasta'. An 'Execute' button is at the bottom.

Galaxy Pasteur

Tools

search tools

Upload Data

into representative sequences

FASTA manipulation

Phylogeny

MAFFT Multiple alignment program for amino acid or nucleotide sequences

alienTrimmer Fast trimming to filter out non-confident nucleotides and alien oligo-nucleotide sequences (adaptors, primers) in both 5' and 3' read ends

condor ConDor: A workflow to detect convergent evolution

FastME Distance-based inference of phylogenetic trees

PastML

Epocs maximum likelihood inference of correlated evolution

Epocs output parser Parsing the result from epocs

MUSCLE Multiple Sequence Comparison by Log- Expectation (Galaxy Version 3.8.1551_0)

! Please provide a value for this option.

Fasta file

No fasta dataset available.

Input file in FASTA format

Run options

Most accurate, maxiters=16

The number of iterations is predefined to allow to select between accuracy or speed options

Find diagonals (-diags)

No

Use diagonal optimization. This can significantly improve speed with closely-related sequences.

Clustering (tree-building) options

UPGMB

Algorithm use to build guide trees for progressive alignment

Output format

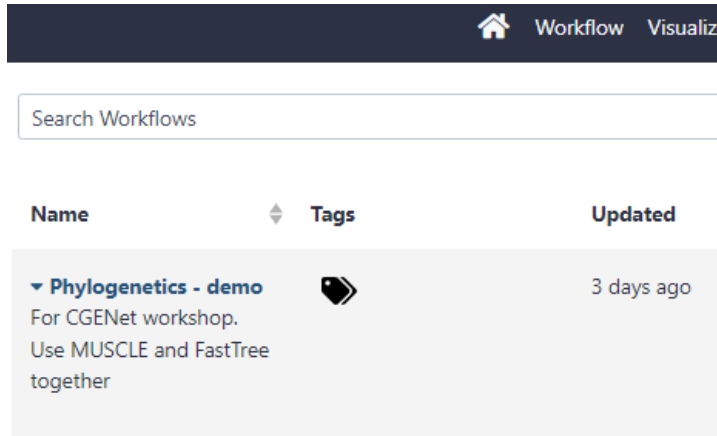
Fasta

Execute

MUSCLE version v3.8.1551

- ❑ 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

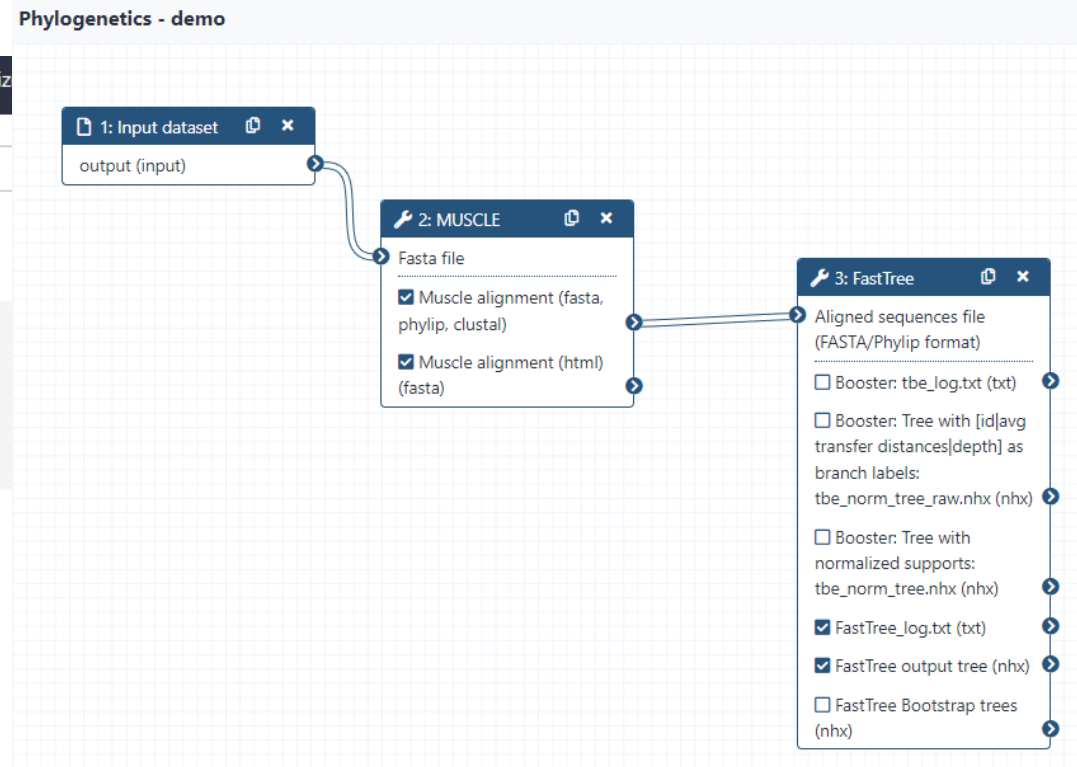


Workflow Visualiz

Search Workflows

Name	Tags	Updated
Phylogenetics - demo		3 days ago
For CGENet workshop. Use MUSCLE and FastTree together		

- ❑ 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
TTCGTGCCGTA CTGACGCGAACGCCAGCAAGTGCTACCGCAAGTCAGGATTCTACCAC
CAT-----GGTGAAAAACAAGGCGGCGGCAGCGGCGGCAGCGGCGGCAGC
GGCAACGAGCCTCGA-----AACTCGTACGCTGACTT
CATAGACAAGAAACAAGCTTTTCGCAAGATACTC-----GTCGCTTC
TATCAACTGGGACGATGTCAGGAACACGCTCAAGGAGGAGTTTATCGGAGAGTGCCGCCA
GAGAA-----ACGGGTCATCATCAGGATATCTGTACGACTA-----
-----CTCCGCAGCCTTAAACCTGACGCTGCGTGCCGCAGAACTGCGCCGCGGTAAC
CGCTCGCAAGCCGAGAACGCTACGATCATACTCGAATTCGATACGGCTAACGTCACCGTA
GG-----
-----TTTTGT
CTGTAATGCTTCGGTTGCGGAAGTCTCGTAC-CGCATCGTCGGCGATACCGC-----
-----TTCTTCGTCGGCGATGCGTACGAACTCTTCTGAAATGT-----
---CTGTTGAGGTGACCTTCTCCGGGATATC-----GGTAGCGC-----
-----AACCTTACATGAGATCCGCATCCTGCACGGATGACTGTTTCG
AAATAGCCACCGAAGCCATCTATCTACAGGCGAACC GTTTACCTCAGCAGGTACAGACA
CTCCCAACGCCACGGCGGAAGCTAC---TACAGCAGAA-----GC
TACTACTCCAGCACCTACCACTCCTGTAC-----CTACCACTCC-----
----CGTACCCTACACGCTGTT-CCAGAAGCGCGTCGCGGAGCGGTGAT-CATCTCTGA
```

History

Unnamed history

8.81 MB 5 9 4

☒ ☐

17 : FastTree output tree

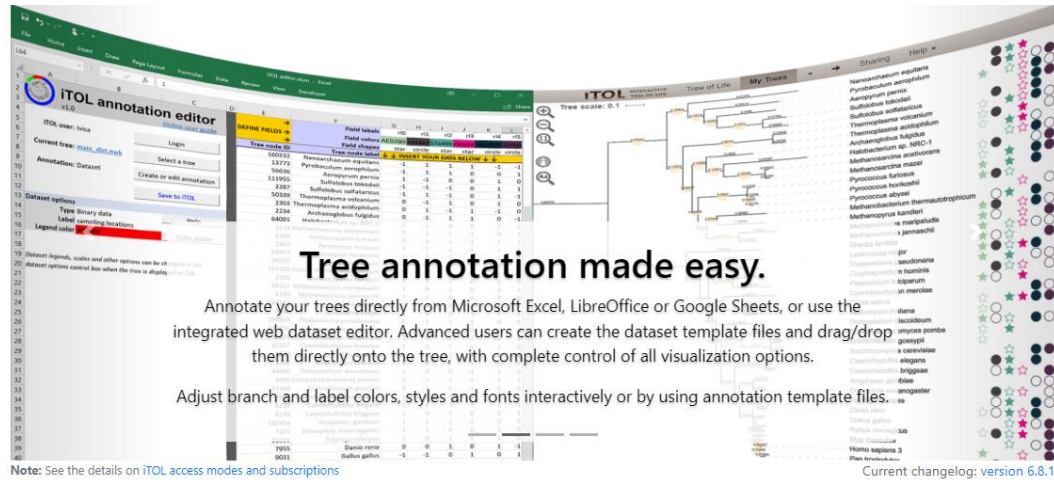
16 : FastTree_log.txt

12 : Muscle alignment (html)

11 : Muscle alignment

10 : Surface-glycoprotein.fasta

Phylogenetic tree viewer



Tree annotation made easy.

Annotate your trees directly from Microsoft Excel, LibreOffice or Google Sheets, or use the integrated web dataset editor. Advanced users can create the dataset template files and drag/drop them directly onto the tree, with complete control of all visualization options.

Adjust branch and label colors, styles and fonts interactively or by using annotation template files.

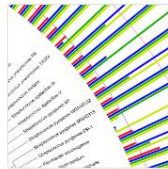
Note: See the details on iTOL access modes and subscriptions

Current changelog: version 6.8.1



Dataset	Tree	Annotations	Colors
1	1000000000	1000000000	1000000000
2	1000000000	1000000000	1000000000
3	1000000000	1000000000	1000000000
4	1000000000	1000000000	1000000000
5	1000000000	1000000000	1000000000

Manage



Annotate



Export



IQ-TREE

Efficient software for phylogenomic inference

Stable release 1.6.12 (August 15, 2019)

[Download v1.6.12 for Windows](#)

Latest release 2.2.2.6 (May 27, 2023)

[Download v2.2.2.6 for Windows](#)

[All Downloads](#)

[Documentation](#)

Let's try analyzing some data



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

