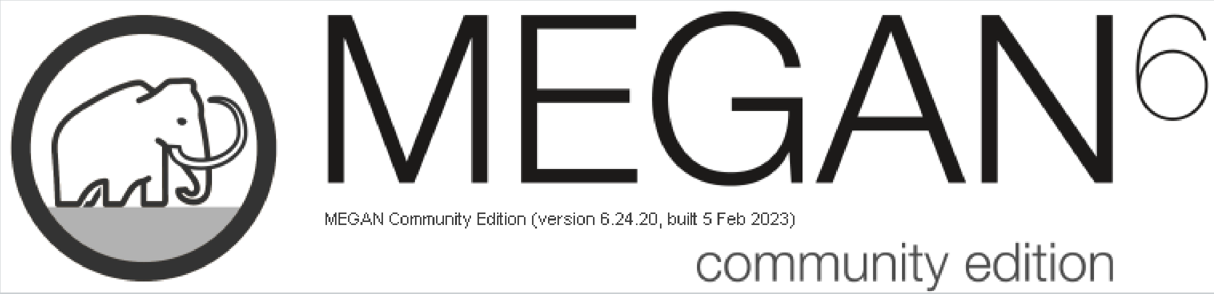
CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN

Bioinformatics Workshop

MEGAN Analysis
3.8.2023
Therese Muzeniek

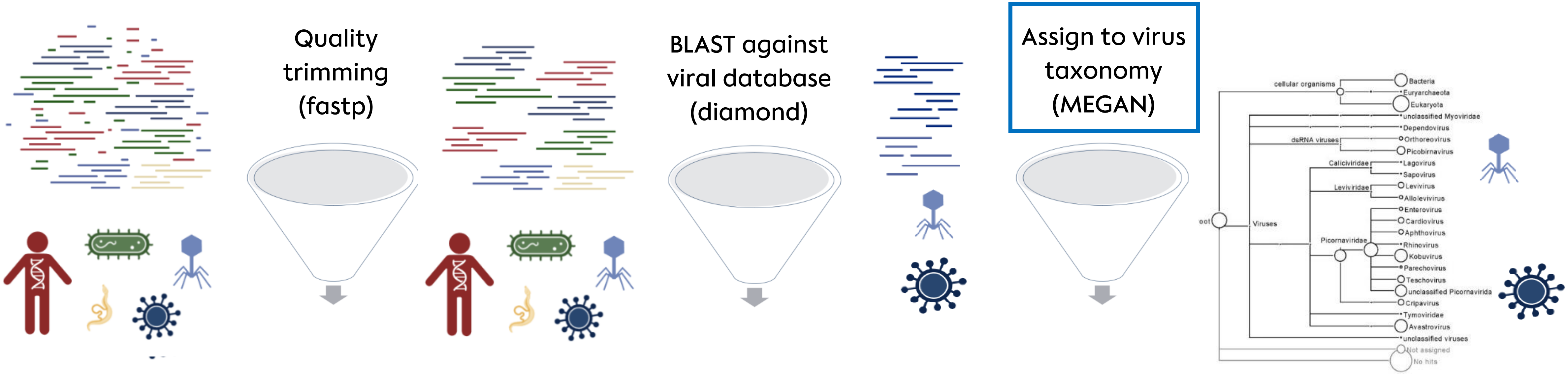
MEGAN Analysis



MEGAN Community Edition (version 6.24.20, built 5 Feb 2023)

Illumina NGS workflow

Overview



Quality trimming (fastp)

BLAST against viral database (diamond)

Assign to virus taxonomy (MEGAN)

➡ Detailed Analysis of viruses of interest in Geneious Prime

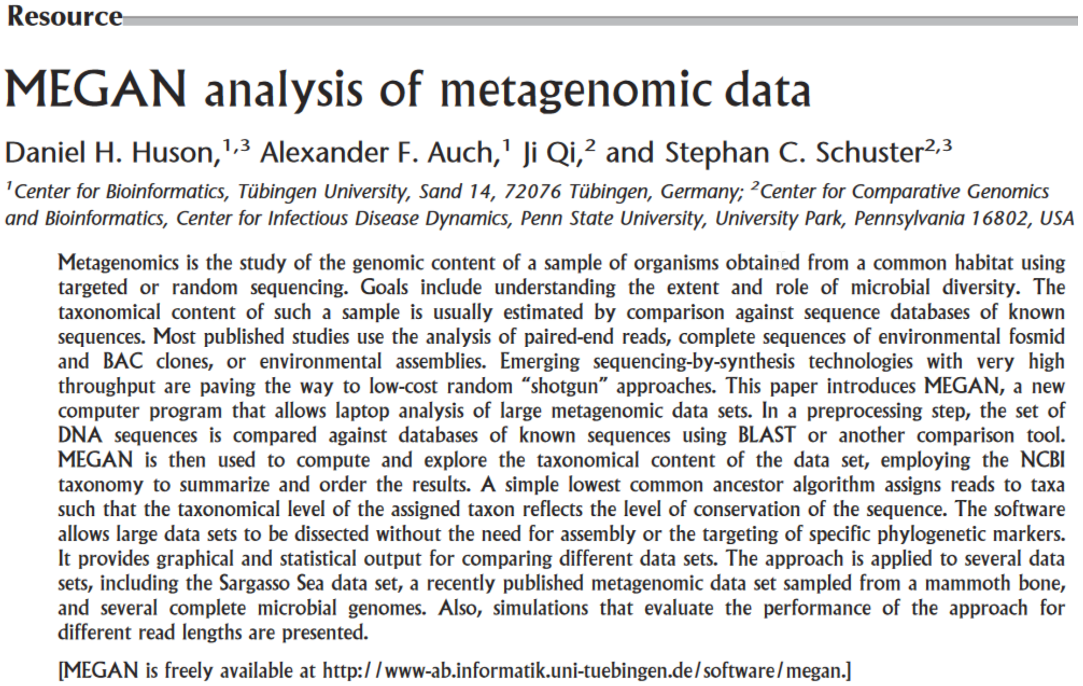
MEGAN6

MEGAN Community Edition (version 6.24.20, built 5 Feb 2023)

community edition

- **ME**ta**G**enome **AN**alyzer
- Visualization tool for diamond analysis data
- Alignment data are visualized in a taxonomic tree

- Input:
  - → .daa file from diamond analysis
  - → NCBI taxonomy database
- Output:
  - → Taxonomic tree
  - → Alignments are assigned to viruses (up to species level)
  - → Reads per species can be further inspected



Resource

MEGAN analysis of metagenomic data

Daniel H. Huson,[1,3] Alexander F. Auch,[1] Ji Qi,[2] and Stephan C. Schuster[2,3]

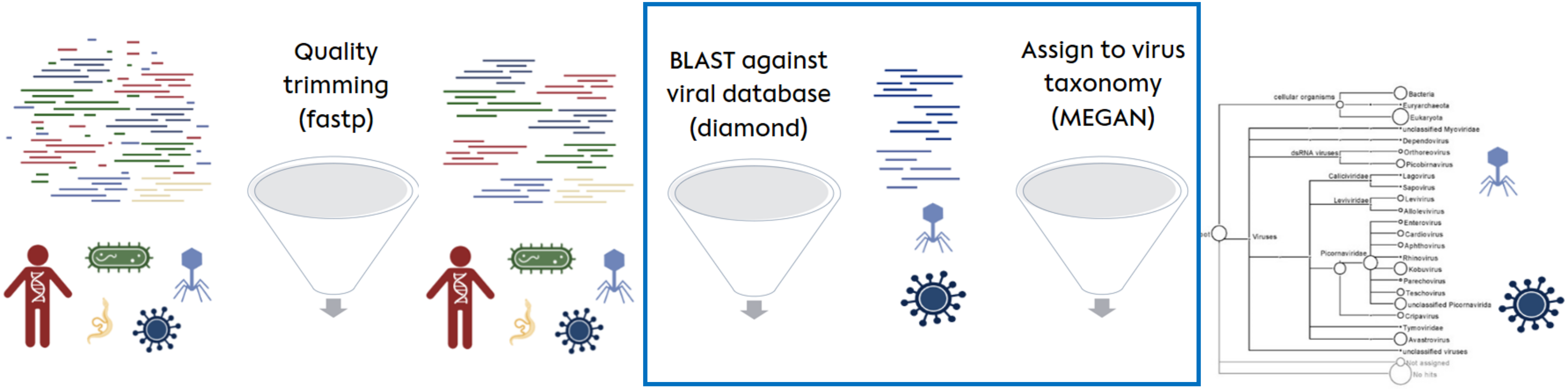
[1]Center for Bioinformatics, Tübingen University, Sand 14, 72076 Tübingen, Germany; [2]Center for Comparative Genomics and Bioinformatics, Center for Infectious Disease Dynamics, Penn State University, University Park, Pennsylvania 16802, USA

Metagenomics is the study of the genomic content of a sample of organisms obtained from a common habitat using targeted or random sequencing. Goals include understanding the extent and role of microbial diversity. The taxonomical content of such a sample is usually estimated by comparison against sequence databases of known sequences. Most published studies use the analysis of paired-end reads, complete sequences of environmental fosmid and BAC clones, or environmental assemblies. Emerging sequencing-by-synthesis technologies with very high throughput are paving the way to low-cost random "shotgun" approaches. This paper introduces MEGAN, a new computer program that allows laptop analysis of large metagenomic data sets. In a preprocessing step, the set of DNA sequences is compared against databases of known sequences using BLAST or another comparison tool. MEGAN is then used to compute and explore the taxonomical content of the data set, employing the NCBI taxonomy to summarize and order the results. A simple lowest common ancestor algorithm assigns reads to taxa such that the taxonomical level of the assigned taxon reflects the level of conservation of the sequence. The software allows large data sets to be dissected without the need for assembly or the targeting of specific phylogenetic markers. It provides graphical and statistical output for comparing different data sets. The approach is applied to several data sets, including the Sargasso Sea data set, a recently published metagenomic data set sampled from a mammoth bone, and several complete microbial genomes. Also, simulations that evaluate the performance of the approach for different read lengths are presented.

[MEGAN is freely available at http://www-ab.informatik.uni-tuebingen.de/software/megan.]

MEGAN Analysis

Preprocessing in the pipeline



Quality trimming (fastp)

BLAST against viral database (diamond)

Assign to virus taxonomy (MEGAN)

- ➢ after diamond BLASTx, the output file (.daa) needs to be meganized
- ➢ Meganize process is part of the pipeline
- ➢ Using the viral protein reference database to parse and analyze the BLAST hits after diamond analysis

MEGAN Analysis

General Tasks

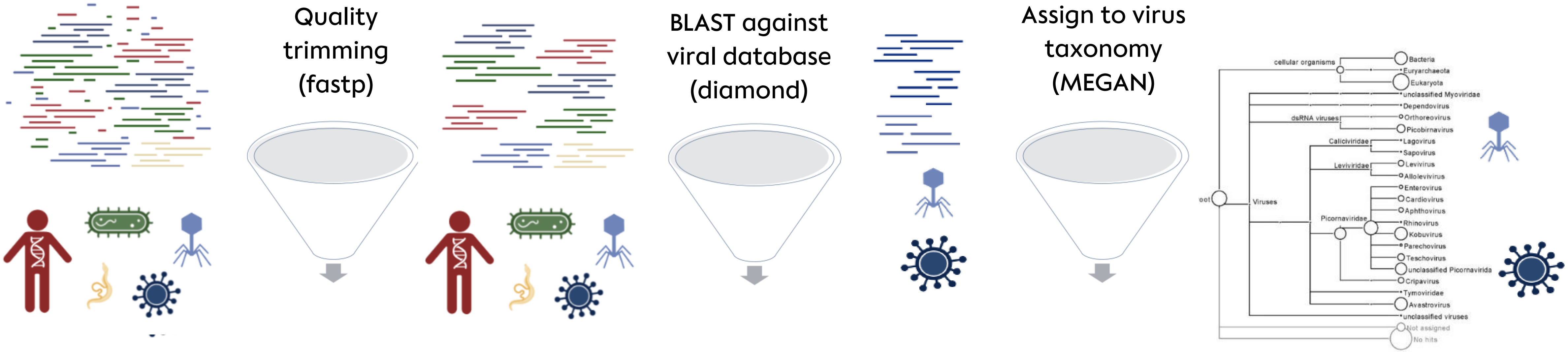
- Open a .daa file in MEGAN
- Collapse and Uncollapse Subtrees
- Display the taxonomic tree on different levels: Order, Family, Genus, Species
    - → Which are orders of interest?
- Find a specific virus species of interest
    - → List summary
    - → List Path
- Find and display the number of assigned reads for a virus species
- Export a summary of counted reads
- Compare different samples in MEGAN

Practical Exercise

Geneious

Illumina NGS workflow

Overview



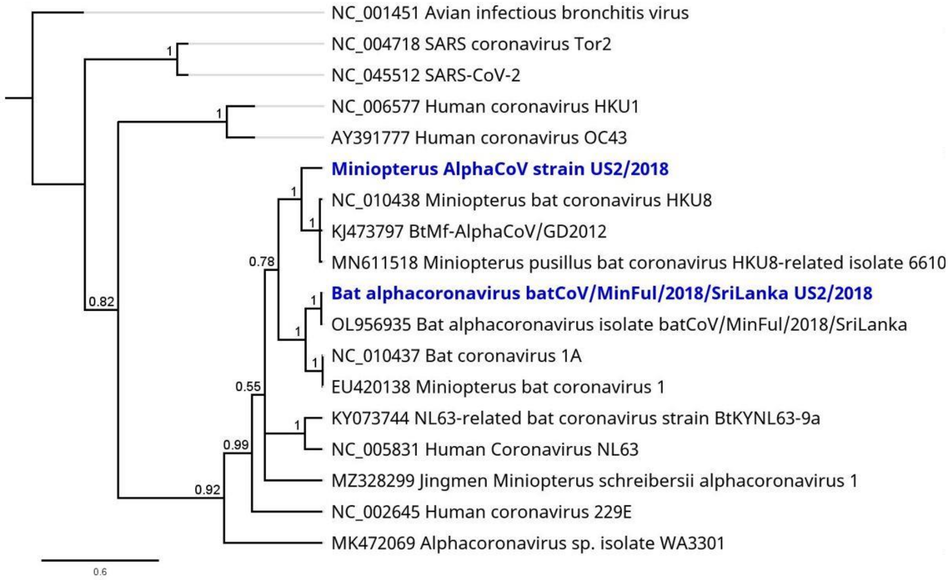
Quality trimming (fastp)

BLAST against viral database (diamond)

Assign to virus taxonomy (MEGAN)

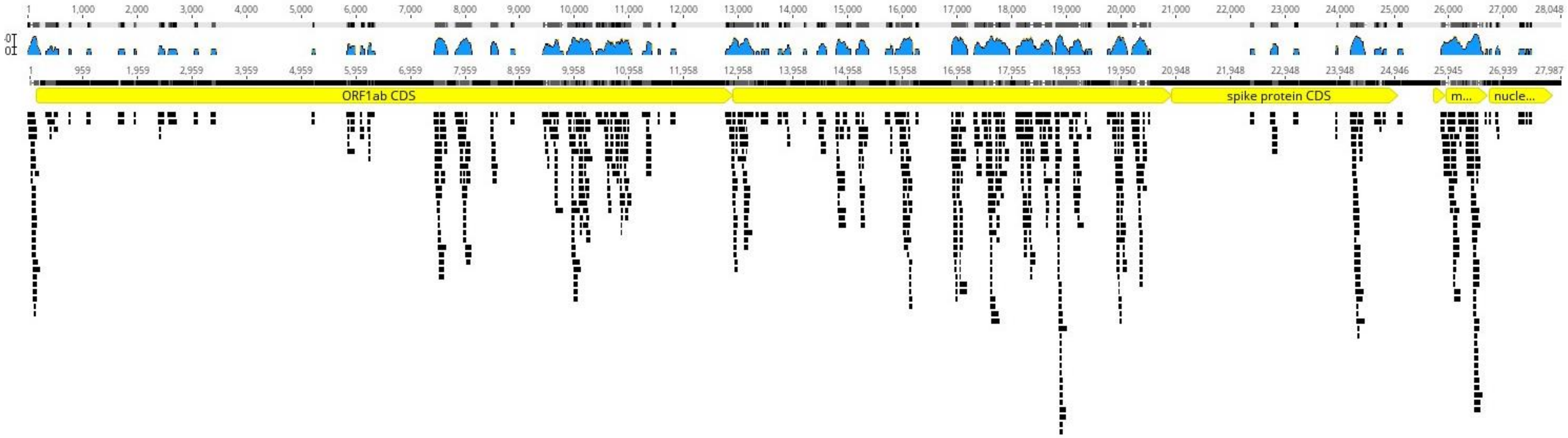
Detailed Analysis of viruses of interest in Geneious Prime

Pipeline tools - geneious prime



NC_001451 Avian infectious bronchitis virus
NC_004718 SARS coronavirus Tor2
NC_045512 SARS-CoV-2
NC_006577 Human coronavirus HKU1
AY391777 Human coronavirus OC43
Miniopterus AlphaCoV strain US2/2018
NC_010438 Miniopterus bat coronavirus HKU8
KJ473797 BtMf-AlphaCoV/GD2012
MN611518 Miniopterus pusillus bat coronavirus HKU8-related isolate 6610
Bat alphacoronavirus batCoV/MinFul/2018/SriLanka US2/2018
OL956935 Bat alphacoronavirus isolate batCoV/MinFul/2018/SriLanka
NC_010437 Bat coronavirus 1A
EU420138 Miniopterus bat coronavirus 1
KY073744 NL63-related bat coronavirus strain BtKYNL63-9a
NC_005831 Human Coronavirus NL63
MZ328299 Jingmen Miniopterus schreibersii alphacoronavirus 1
NC_002645 Human coronavirus 229E
MK472069 Alphacoronavirus sp. isolate WA3301
1
1
1
1
1
1
1
1
0.78
0.82
0.55
0.99
0.92
0.6

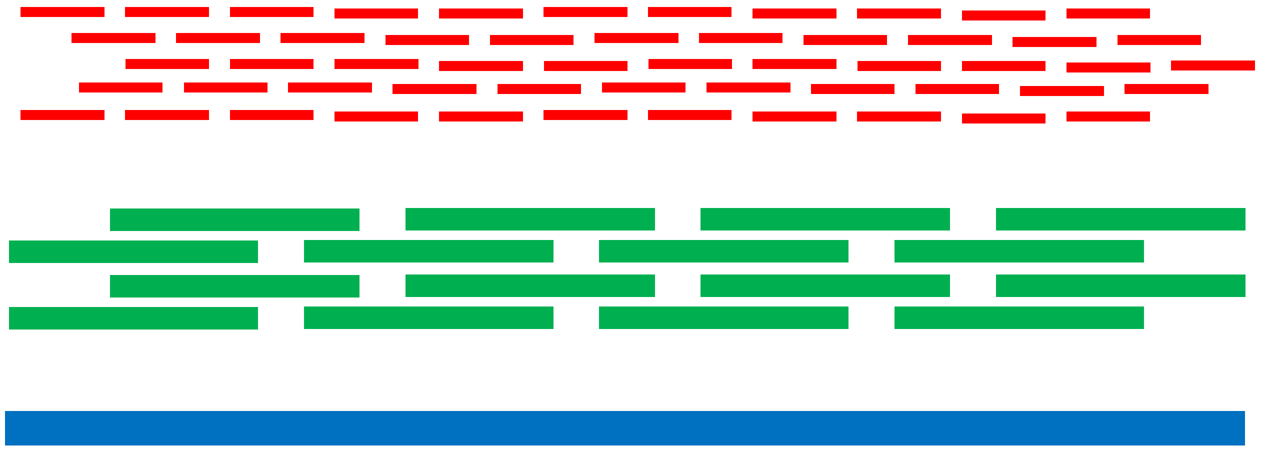
- Comprehensive platform for a wide range of data analysis tasks:
  - → Validation of results
  - → Alignment of reads
  - → Assembly of reads to a reference genome
  - → Annotation of genes
  - → Phylogenetic analyses



Geneious Prime

General Tasks

- Create a new folder and add data (extracted reads) from MEGAN
- Optional: de novo assembly of reads to create longer contigs



NGS reads

Longer contigs after de novo assembly

Full Genome

Geneious Prime

General Tasks

- BLAST search of reads (or de novo assembled contigs) in Geneious
  - → Different options and settings
- Download reference sequences
- Map to reference
  - → Different mapper and settings are available

Geneious Prime

Practical Exercises

- Compare NGS data sets from different samples (f.e. human SARS vs human Polio example data) using MEGAN

  → Do you find the typical contaminants (f.e. Bacteriophages, hits mapping large viral genomes etc…)?

- Compare NGS data from different sample types (f.e. human example data vs livestock data)

  → What are the differences, what is same?

- Analyse a data set using different in the pipeline (f.e. change the sensitivity settings for diamond) and compare the settings in MEGAN

  → What are the differences (total reads, assigned reads etc.)

- Extract reads of interest from the previous exercises and analyse it in Geneious

- Design primers for a retesting PCR assay