CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN

Bioinformatics Workshop

Udo Gieraths, Therese Muzeniek
31.7. – 4.8.2023
KCCR

Contents Morning Session

1. Introduction & Agenda for the week
2. From NGS (Lab work) to data analysis (Computer work)
3. Executing the analysis pipeline with example data

Introduction and Agenda

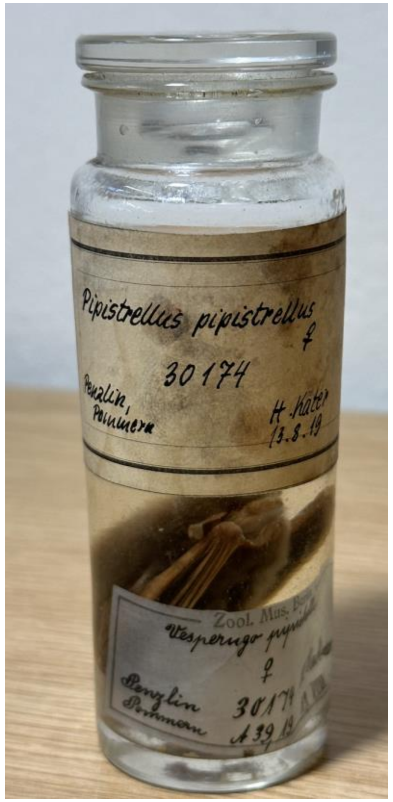
Introduction Udo Gieraths

- Background in
  - Computer Science (Bachelor)
  - Bioinformatics (Master)
  - Data Science (Booking.com, Soundcloud)
- PhD studies since ~3 years at Charite
- Focus on virus discovery in ancient samples

Introduction Udo Gieraths

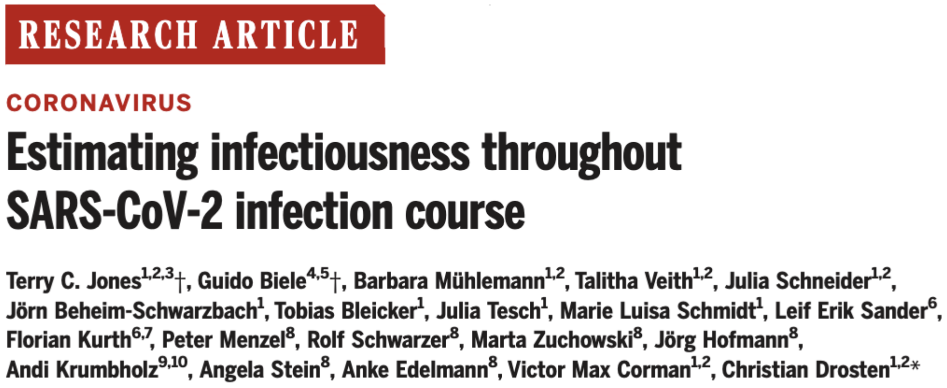
- My PhD topic
  - Ancient viral RNA/DNA genome discovery

- Writing NGS pipelines
  - to detect just traces of viral RNA/DNA
  - to extract as many matching reads as possible
- Reconstruct the ancient genome



Introduction Udo Gieraths

- Group of Terry Jones
  - 3 PhD students (2 Bioinformatics, 1 Bioinfo/Lab)
  - 1 Postdoc (Bioinformatics)
  - 1 Lab technician
- We handle most NGS processing of the virology department

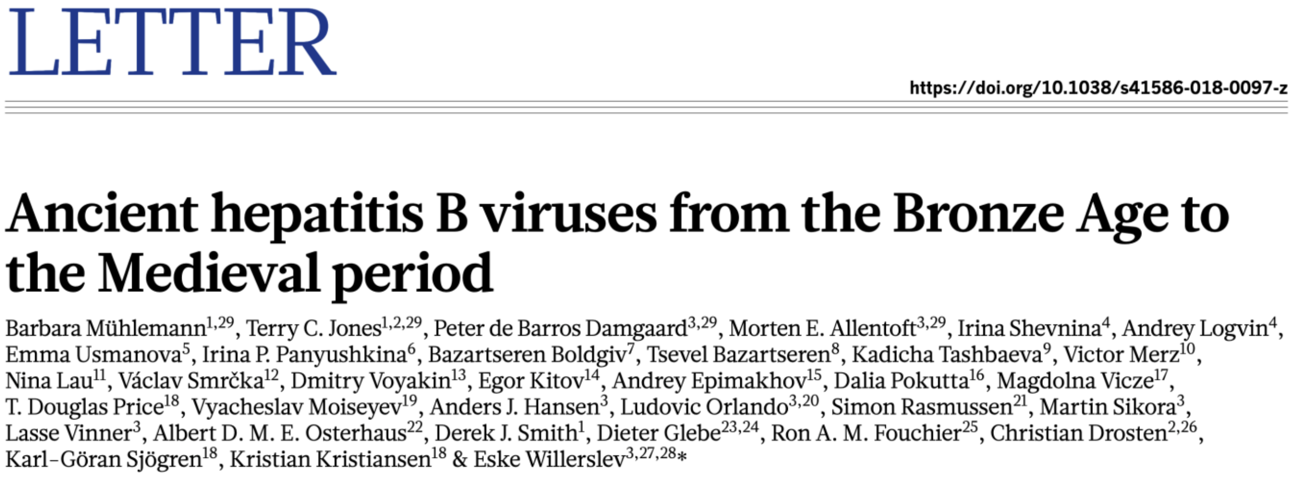


RESEARCH ARTICLE

CORONAVIRUS

Estimating infectiousness throughout SARS-CoV-2 infection course

Terry C. Jones[1,2,3]†, Guido Biele[4,5]†, Barbara Mühlemann[1,2], Talitha Veith[1,2], Julia Schneider[1,2], Jörn Beheim-Schwarzbach[1], Tobias Bleicker[1], Julia Tesch[1], Marie Luisa Schmidt[1], Leif Erik Sander[6], Florian Kurth[6,7], Peter Menzel[8], Rolf Schwarzer[8], Marta Zuchowski[8], Jörg Hofmann[8], Andi Krumbholz[9,10], Angela Stein[8], Anke Edelmann[8], Victor Max Corman[1,2], Christian Drosten[1,2]*



LETTER

https://doi.org/10.1038/s41586-018-0097-z

Ancient hepatitis B viruses from the Bronze Age to the Medieval period

Barbara Mühlemann[1,29], Terry C. Jones[1,2,29], Peter de Barros Damgaard[3,29], Morten E. Allentoft[3,29], Irina Shevnina[4], Andrey Logvin[4], Emma Usmanova[5], Irina P. Panyushkina[6], Bazartseren Boldgiv[7], Tsevel Bazartseren[8], Kadicha Tashbaeva[9], Victor Merz[10], Nina Lau[11], Václav Smrčka[12], Dmitry Voyakin[13], Egor Kitov[14], Andrey Epimakhov[15], Dalia Pokutta[16], Magdolna Vicze[17], T. Douglas Price[18], Vyacheslav Moiseyev[19], Anders J. Hansen[3], Ludovic Orlando[3,20], Simon Rasmussen[21], Martin Sikora[3], Lasse Vinner[3], Albert D. M. E. Osterhaus[22], Derek J. Smith[1], Dieter Glebe[23,24], Ron A. M. Fouchier[25], Christian Drosten[2,26], Karl-Göran Sjögren[18], Kristian Kristiansen[18] & Eske Willerslev[3,27,28]*

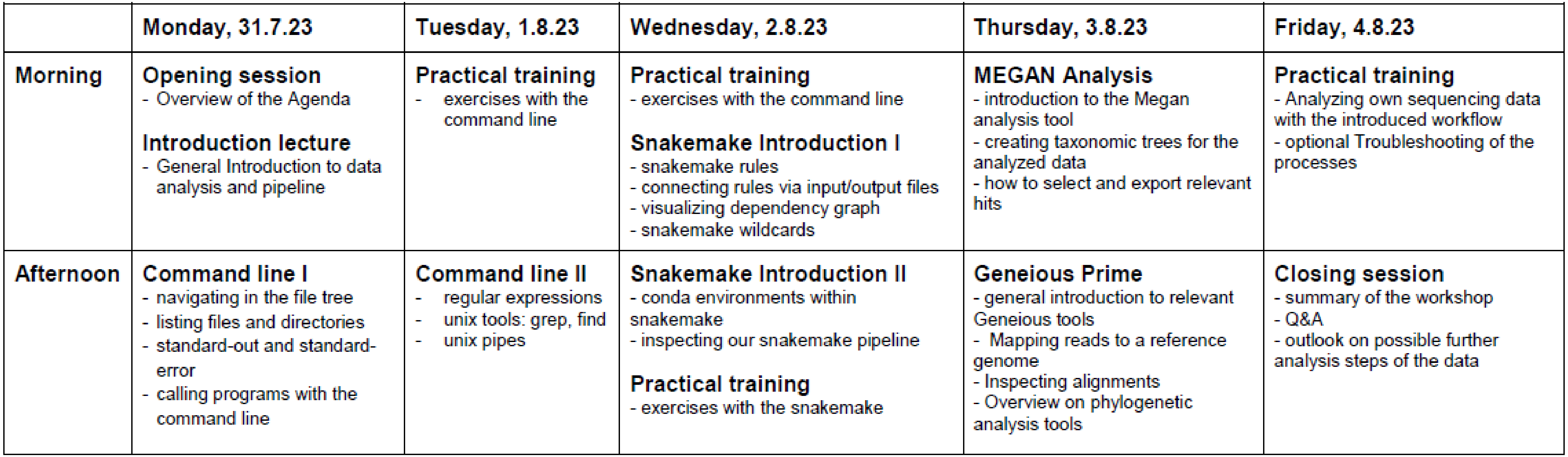
Introduction Therese Muzeniek

- Bachelor and Master of Science in Biotechnology
- Doctor of Natural Sciences (special field Biology)

- Focus on virus discovery in bat samples
  - ➢ metagenomic NGS and comparative virome analysis
  - ➢ Full genome assembly of viruses from NGS data
  - ➢ Phylogenetic analysis of novel virus strains

Introduction Workshop Participants

Workshop Agenda

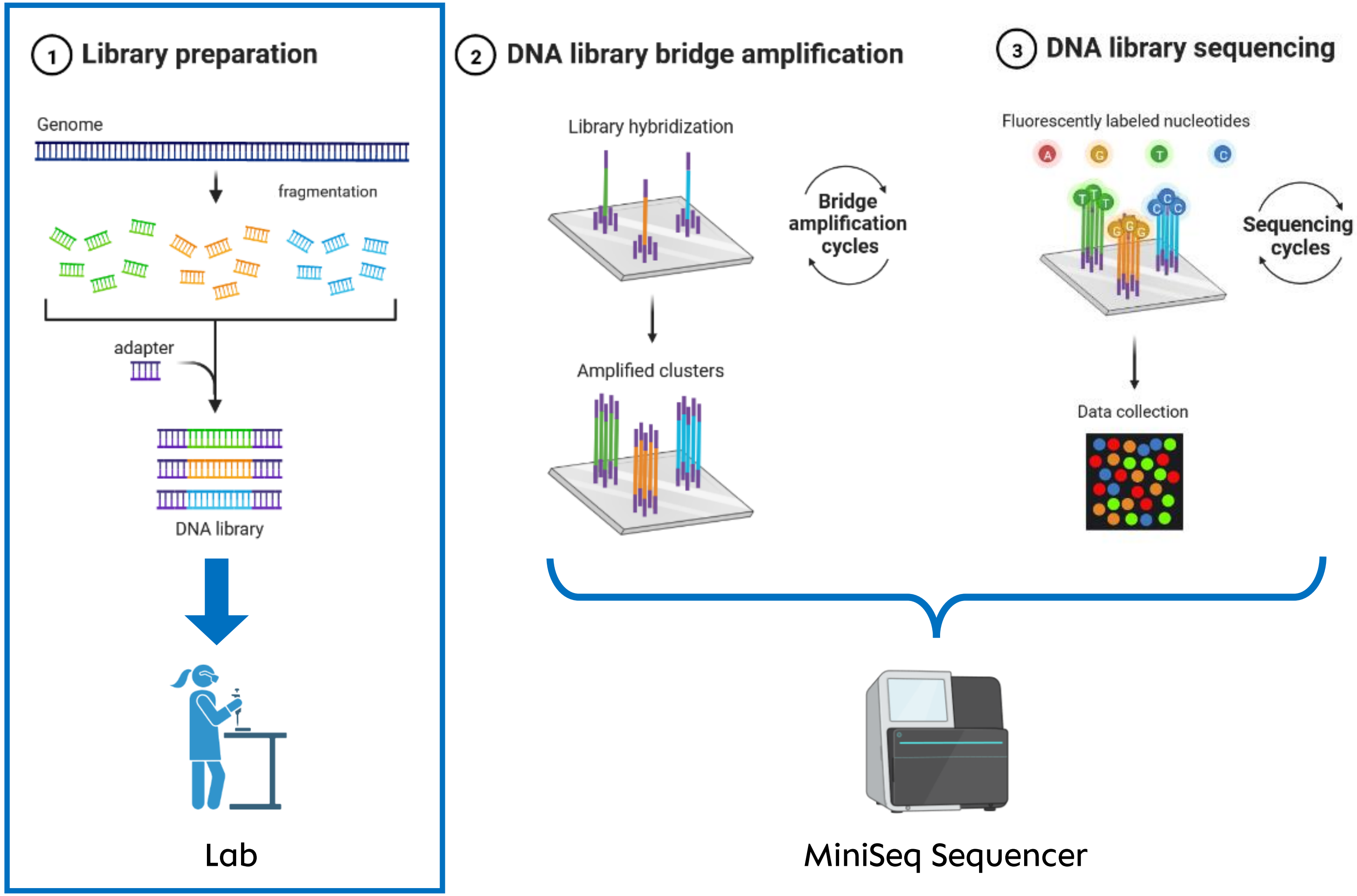


	Monday, 31.7.23	Tuesday, 1.8.23	Wednesday, 2.8.23	Thursday, 3.8.23	Friday, 4.8.23
Morning	**Opening session** - Overview of the Agenda **Introduction lecture** - General Introduction to data analysis and pipeline	**Practical training** - exercises with the command line	**Practical training** - exercises with the command line **Snakemake Introduction I** - snakemake rules - connecting rules via input/output files - visualizing dependency graph - snakemake wildcards	**MEGAN Analysis** - introduction to the Megan analysis tool - creating taxonomic trees for the analyzed data - how to select and export relevant hits	**Practical training** - Analyzing own sequencing data with the introduced workflow - optional Troubleshooting of the processes
Afternoon	**Command line I** - navigating in the file tree - listing files and directories - standard-out and standard-error - calling programs with the command line	**Command line II** - regular expressions - unix tools: grep, find - unix pipes	**Snakemake Introduction II** - conda environments within snakemake - inspecting our snakemake pipeline **Practical training** - exercises with the snakemake	**Geneious Prime** - general introduction to relevant Geneious tools - Mapping reads to a reference genome - Inspecting alignments - Overview on phylogenetic analysis tools	**Closing session** - summary of the workshop - Q&A - outlook on possible further analysis steps of the data

From NGS to Data analysis

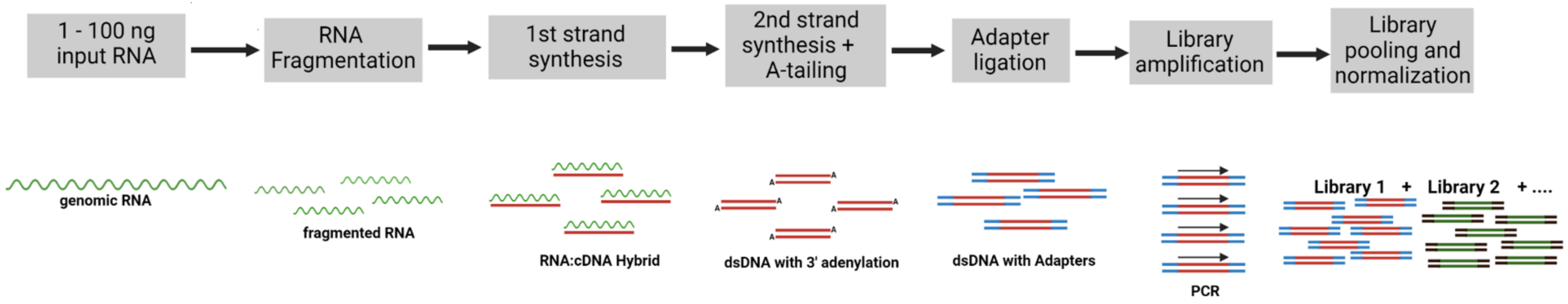
Illumina NGS

General Principle



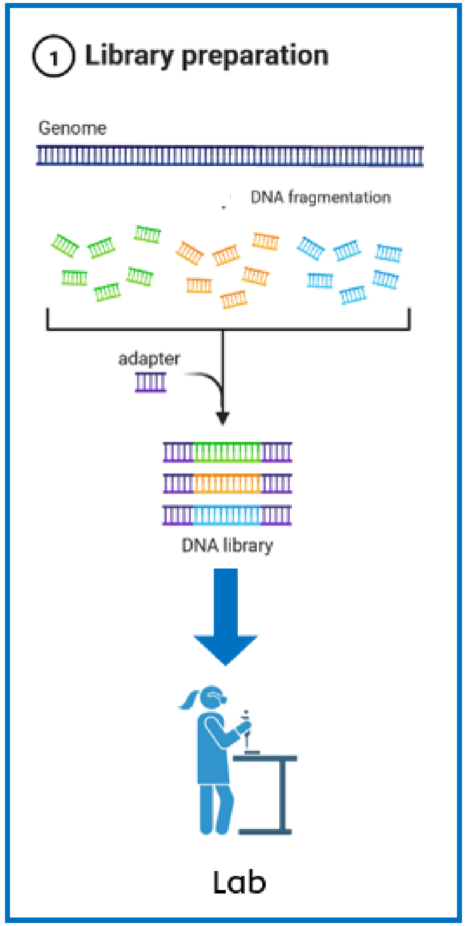
Library Preparation

KAPA RNA HyperPrep Kit (Roche)



Library Preparation

KAPA RNA HyperPrep Kit (Roche)



Library Preparation of the whole RNA (+DNA) content in a sample

→ Allows for an unbiased (untargeted/ metagenomic) sequencing

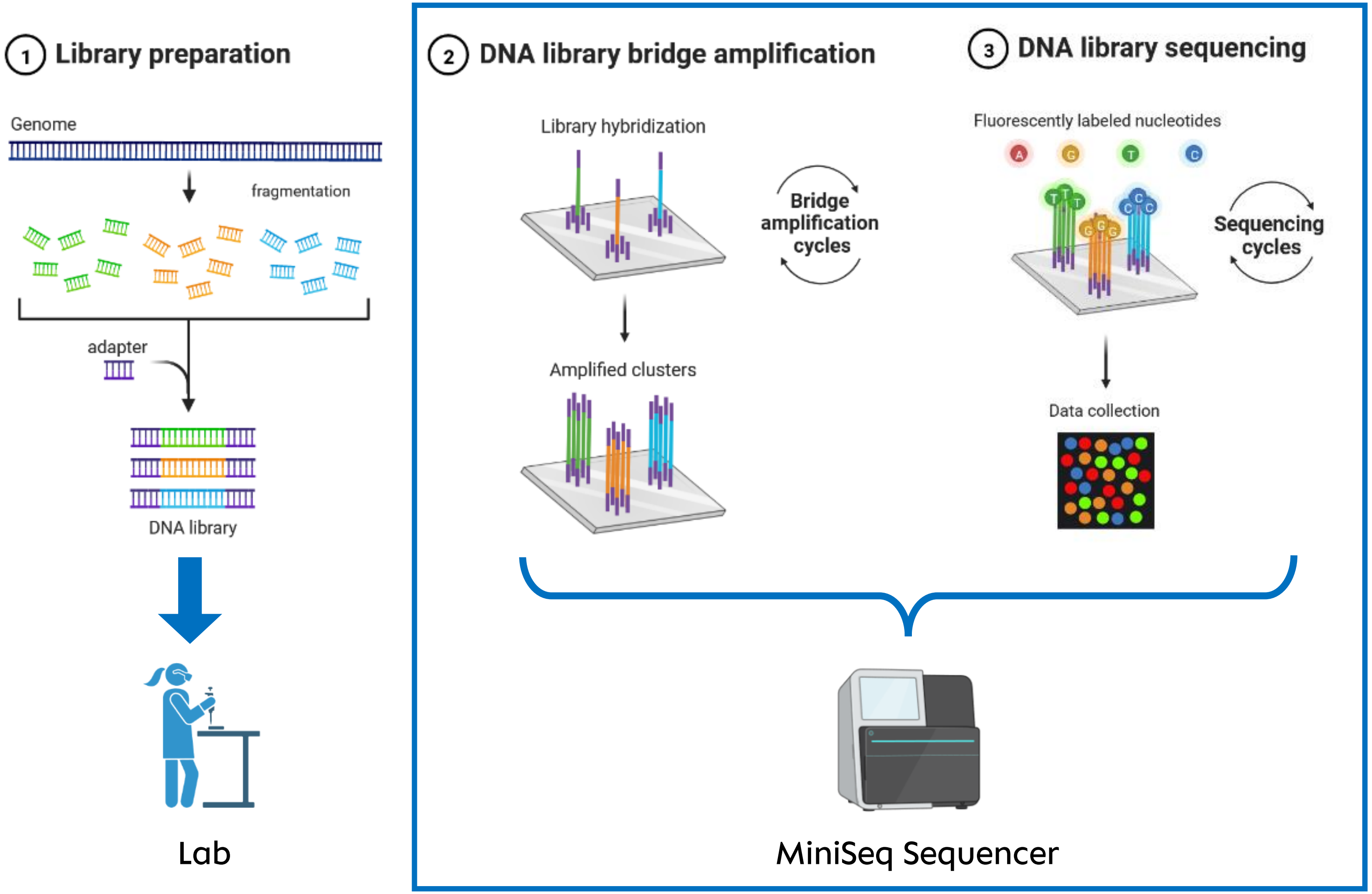
→ Including Host, Bacterial, Viral, Parasite RNA / DNA material

Challenge: How to find the (viral) reads that are of interest?

→ **Specific bioinformatics pipeline for virus discovery**

Illumina NGS

General Principle



1 Library preparation

Genome

fragmentation

adapter

DNA library

Lab

2 DNA library bridge amplification

Library hybridization

Bridge amplification cycles

Amplified clusters

3 DNA library sequencing

Fluorescently labeled nucleotides

A G T C

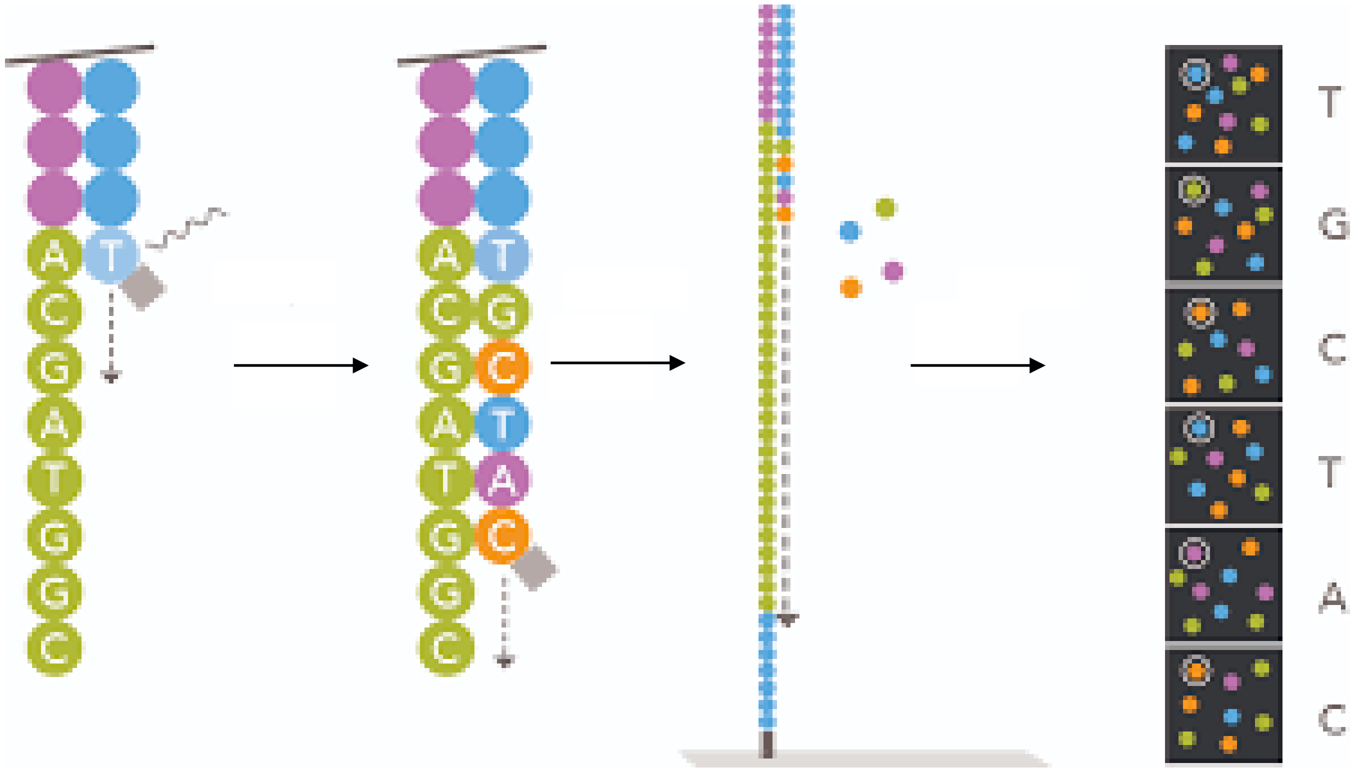
Sequencing cycles

Data collection

MiniSeq Sequencer

Illumina NGS

Sequencing by Synthesis



Basecalling

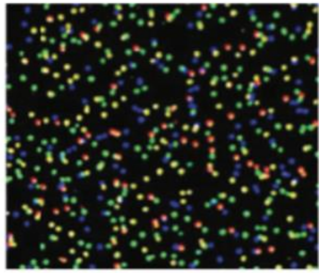
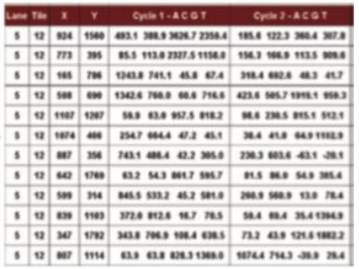
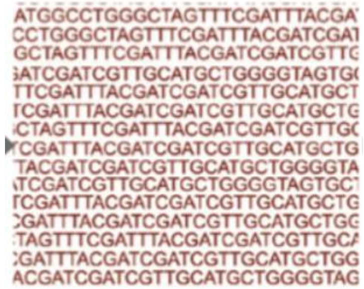


Image Analysis



Lane	Tile	X	Y	Cycle 1 - A C G T	Cycle 2 - A C G T
5	12	924	1560	493.1 388.9 3626.7 2358.4	185.6 122.3 360.4 307.8
5	12	773	395	85.5 113.0 2327.5 1158.0	156.3 166.9 113.5 909.6
5	12	165	786	1243.8 741.1 45.8 67.4	318.4 692.6 48.3 41.7
5	12	588	690	1342.6 760.0 60.6 716.6	423.6 505.7 1010.1 958.3
5	12	1107	1207	58.9 63.0 957.5 818.2	98.6 230.5 815.1 512.1
5	12	1074	408	254.7 664.4 47.2 45.1	38.4 41.8 64.9 1102.9
5	12	887	356	743.1 488.4 42.2 305.0	230.3 603.6 -43.1 -28.1
5	12	642	1760	63.2 54.3 861.7 595.7	81.5 86.0 54.9 385.4
5	12	589	314	845.5 533.2 45.2 581.0	280.9 560.9 13.0 78.4
5	12	839	1103	372.8 812.6 16.7 78.5	58.4 68.4 35.4 1394.9
5	12	347	1792	343.8 796.9 108.4 638.5	73.2 43.9 121.8 1882.2
5	12	807	1114	63.9 63.8 828.3 1360.0	1074.4 714.3 -38.9 29.4

Base Calling



```
ATGGCCTGGGCTAGTTTCGATTTACGA
CCTGGGCTAGTTTCGATTTACGATCGAT
GCTAGTTTCGATTTACGATCGATCGTTG
GATCGATCGTTGCATGCTGGGGTAGTG
TTCGATTTACGATCGATCGTTGCATGCT
TCGATTTACGATCGATCGTTGCATGCTG
CTAGTTTCGATTTACGATCGATCGTTG
TCGATTTACGATCGATCGTTGCATGCTG
TACGATCGATCGTTGCATGCTGGGGTA
TCGATCGTTGCATGCTGGGGTAGTGCT
TCGATTTACGATCGATCGTTGCATGCTG
CGATTTACGATCGATCGTTGCATGCTGG
TAGTTTCGATTTACGATCGATCGTTGCA
CGATTTACGATCGATCGTTGCATGCTGG
ACGATCGATCGTTGCATGCTGGGGTAG
```

Output data = .bcl format

Illumina NGS

Output data and conversion

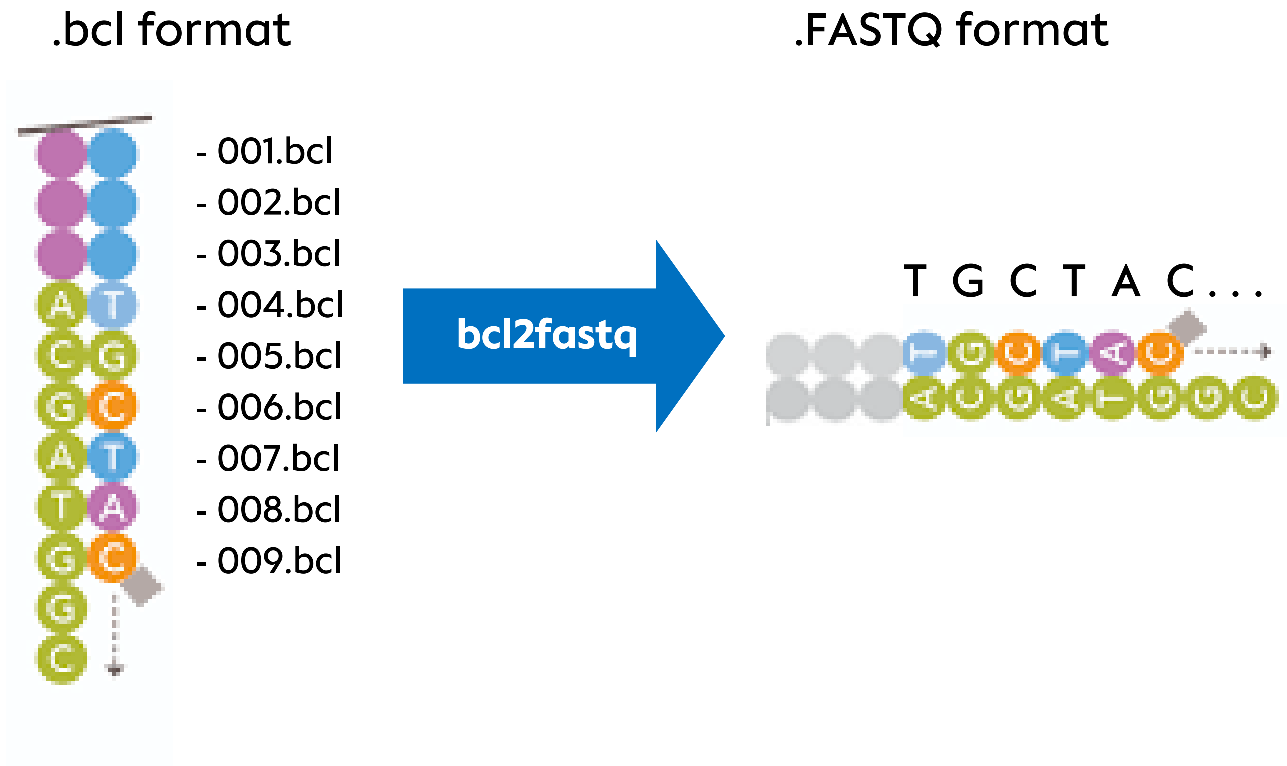
- MiniSeq Output: .bcl files

➢ **Per-cycle** basecall files

➢ Nucleotide information for all clusters per cycle

➢ For data analysis we need **per-read** FASTQ file

➢bcl2fastq tool for conversion + demultiplexing



.bcl format

- 001.bcl
- 002.bcl
- 003.bcl
- 004.bcl
- 005.bcl
- 006.bcl
- 007.bcl
- 008.bcl
- 009.bcl

bcl2fastq

.FASTQ format

T G C T A C . . .

FASTQ file:

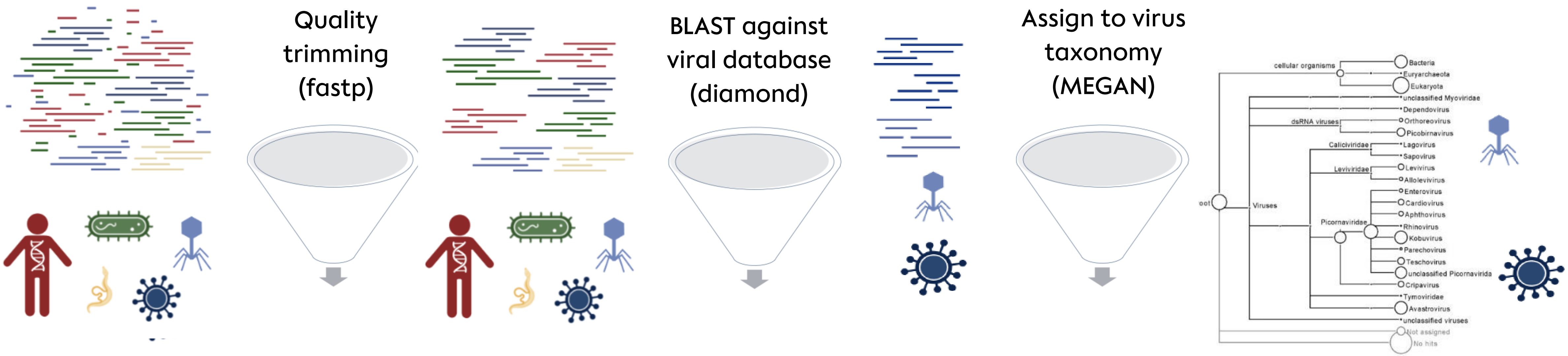
- **Run information**
- **Raw read sequence**
- **Sequencing quality**

```
@MN02032:2:000H5KF7K:1:11102:15995:1048 1:N:0:GTAACATC+AATCGCTG
CCCCTGACCCCTTCGTTTGCNGCGAAGTGCGCNNN
+
FFFFFFAFFFFFFFFFFFFF#FFFFFFFFFFF###
```

FASTQ = Input for Data analysis pipeline

NGS data analysis

Focus virus discovery



Quality trimming (fastp)

BLAST against viral database (diamond)

Assign to virus taxonomy (MEGAN)

➡ Detailed Analysis of viruses of interest in Geneious Prime

Practical Exercise

Executing the analysis pipeline with example data

- Open the command line interface

- Activate Snakemake environment by typing:

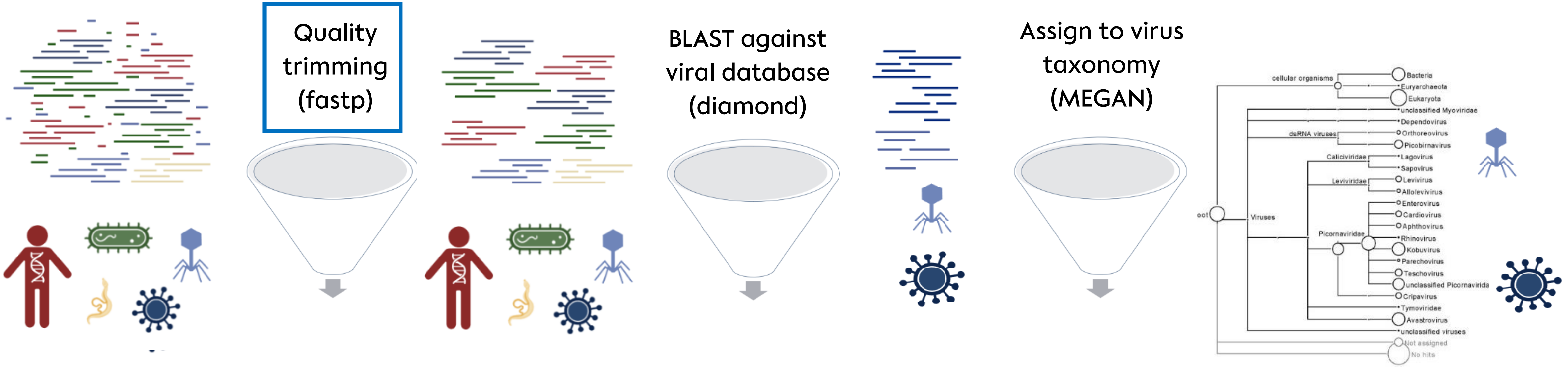
```
conda activate snakemake
```

- Execute the Pipeline by typing:

```
snakemake –s Snakefile_simplified.smk --use-conda --cores 10
```

Meanwhile…

Introduction of the tools used in the pipeline



Quality trimming (fastp)

BLAST against viral database (diamond)

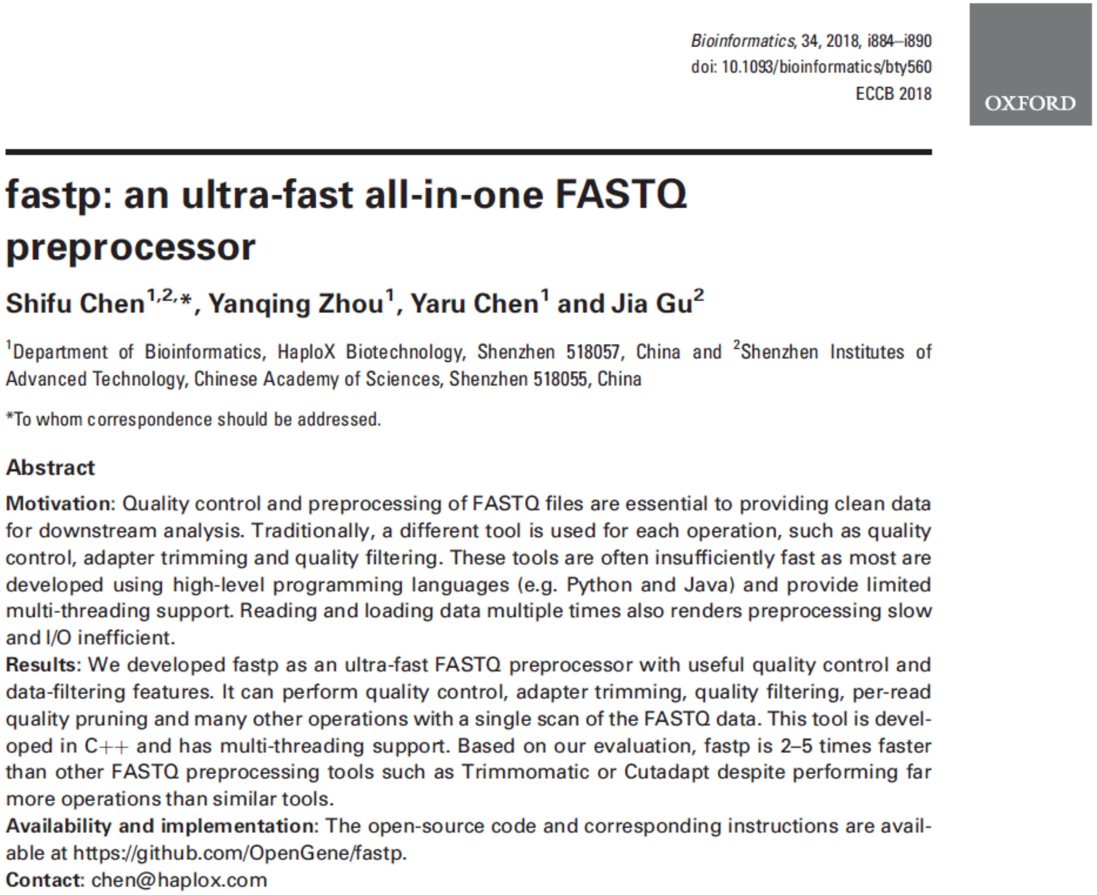
Assign to virus taxonomy (MEGAN)

→ Detailed Analysis of viruses of interest in Geneious Prime

Pipeline tools - Fastp

- Pre-processing tool for FASTQ data
- Fast and efficient
- Processing of Paired-End inputs
- Includes quality control and data filtering
  - → Removal of low quality reads
  - → Removal of reads with too many N
  - → Removal of short reads
  - → Only good quality reads „survive“

- Report is generated as .html document



Bioinformatics, 34, 2018, i884–i890
doi: 10.1093/bioinformatics/bty560
ECCB 2018

OXFORD

fastp: an ultra-fast all-in-one FASTQ preprocessor

Shifu Chen[1,2,*], Yanqing Zhou[1], Yaru Chen[1] and Jia Gu[2]

[1]Department of Bioinformatics, HaploX Biotechnology, Shenzhen 518057, China and [2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China

*To whom correspondence should be addressed.

Abstract

Motivation: Quality control and preprocessing of FASTQ files are essential to providing clean data for downstream analysis. Traditionally, a different tool is used for each operation, such as quality control, adapter trimming and quality filtering. These tools are often insufficiently fast as most are developed using high-level programming languages (e.g. Python and Java) and provide limited multi-threading support. Reading and loading data multiple times also renders preprocessing slow and I/O inefficient.

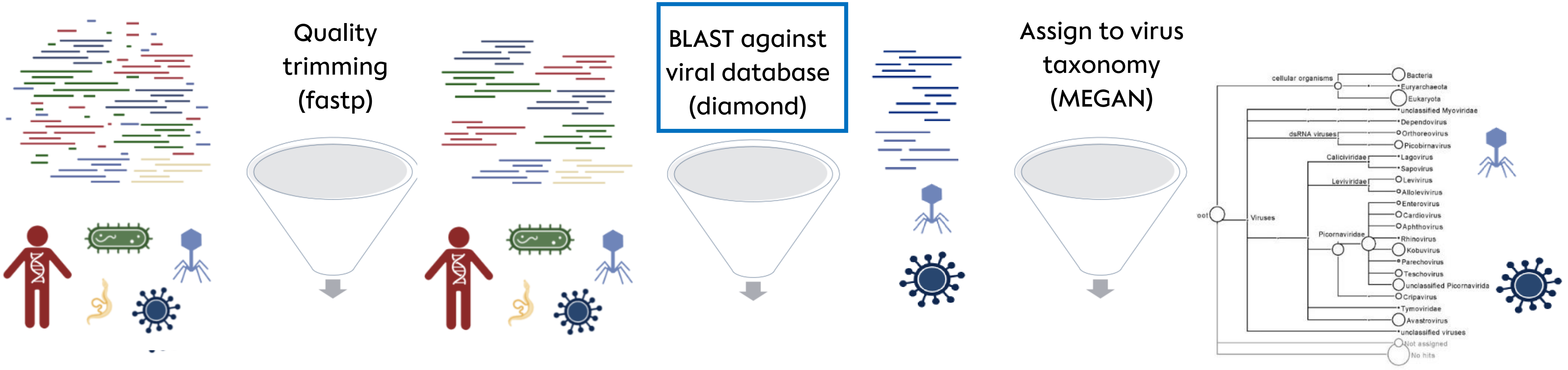
Results: We developed fastp as an ultra-fast FASTQ preprocessor with useful quality control and data-filtering features. It can perform quality control, adapter trimming, quality filtering, per-read quality pruning and many other operations with a single scan of the FASTQ data. This tool is developed in C++ and has multi-threading support. Based on our evaluation, fastp is 2–5 times faster than other FASTQ preprocessing tools such as Trimmomatic or Cutadapt despite performing far more operations than similar tools.

Availability and implementation: The open-source code and corresponding instructions are available at https://github.com/OpenGene/fastp.

Contact: chen@haplox.com

Meanwhile…

Introduction of the tools used in the pipeline



Quality trimming (fastp)

BLAST against viral database (diamond)

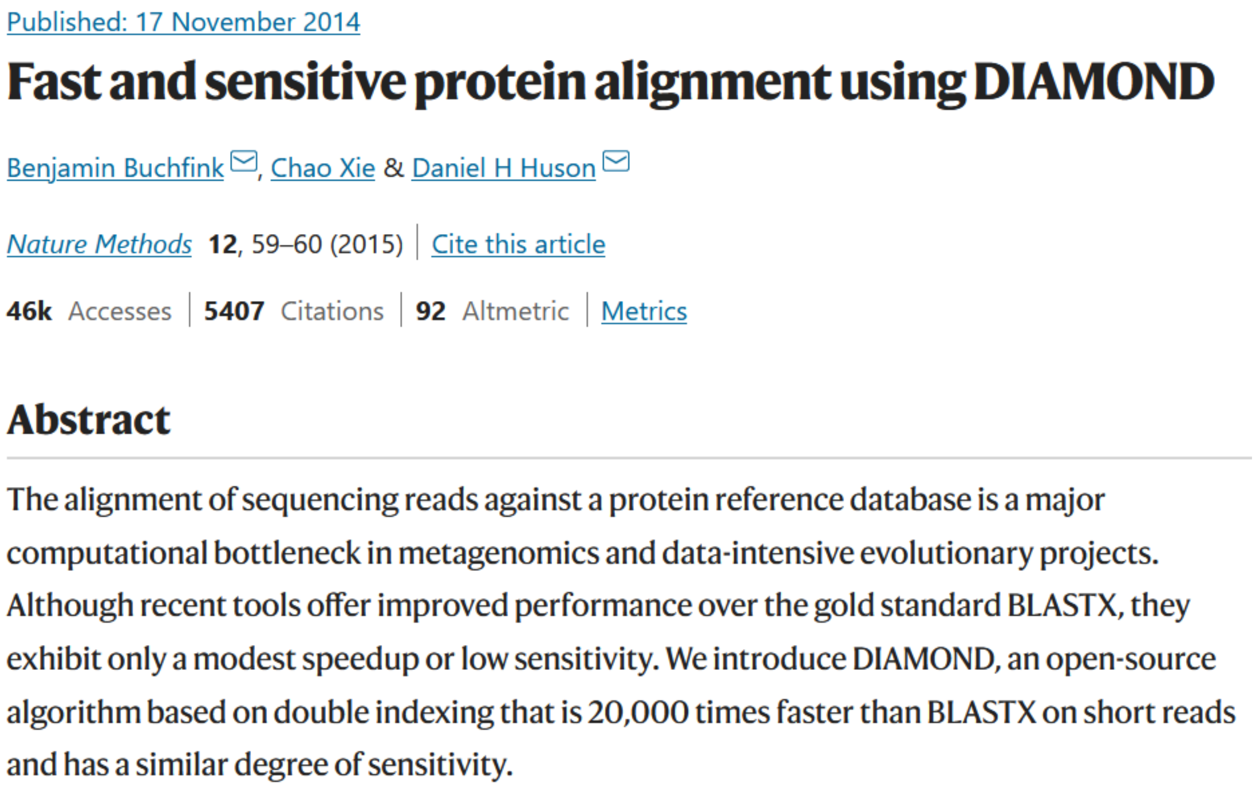
Assign to virus taxonomy (MEGAN)

➡ Detailed Analysis of viruses of interest in Geneious Prime

Pipeline tools - diamond

- “Distributed Alignment of INformative DNA Sequences”
- Sequence aligner for protein and translated DNA
- Up to x20,000 speed of BLAST search
  - → Suitable for NGS data
  - → High-Throughput protein alignment possible

- Aligns DNA reads to a protein reference database
- Input:
  - → FASTQ file after quality trimming
  - → Protein reference database
- Output:
  - → .daa file with alignment results



Published: 17 November 2014

Fast and sensitive protein alignment using DIAMOND

Benjamin Buchfink, Chao Xie & Daniel H Huson

Nature Methods **12**, 59–60 (2015) | Cite this article

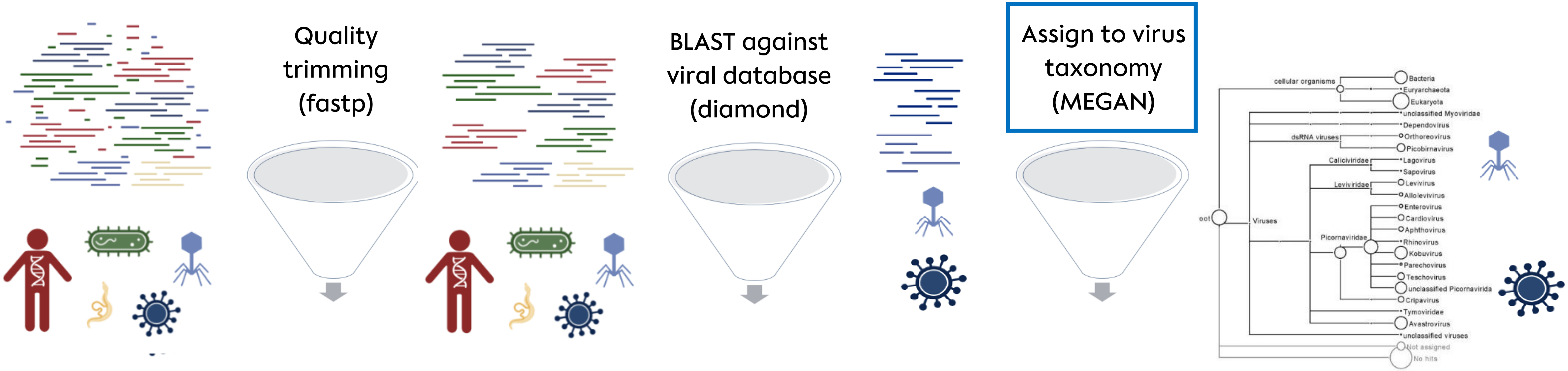
46k Accesses | **5407** Citations | **92** Altmetric | Metrics

Abstract

The alignment of sequencing reads against a protein reference database is a major computational bottleneck in metagenomics and data-intensive evolutionary projects. Although recent tools offer improved performance over the gold standard BLASTX, they exhibit only a modest speedup or low sensitivity. We introduce DIAMOND, an open-source algorithm based on double indexing that is 20,000 times faster than BLASTX on short reads and has a similar degree of sensitivity.

Meanwhile…

Introduction of the tools used in the pipeline



Quality trimming (fastp)

BLAST against viral database (diamond)

Assign to virus taxonomy (MEGAN)

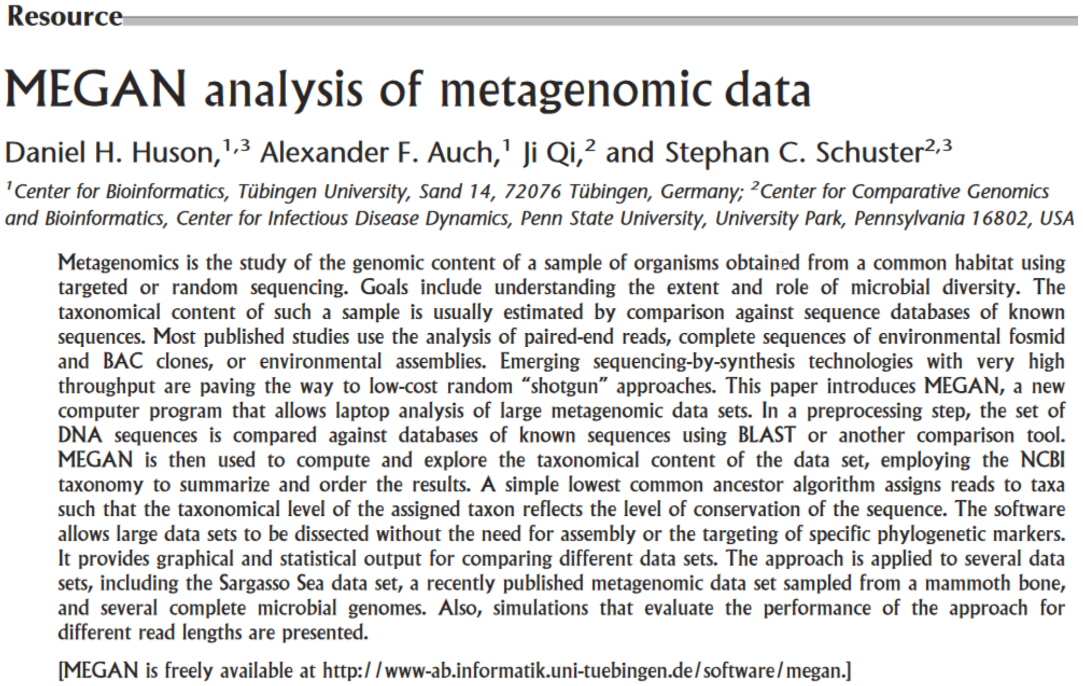
➡ Detailed Analysis of viruses of interest in Geneious Prime

Pipeline tools - 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
  (MEGAN or Geneious)



Resource

MEGAN analysis of metagenomic data

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

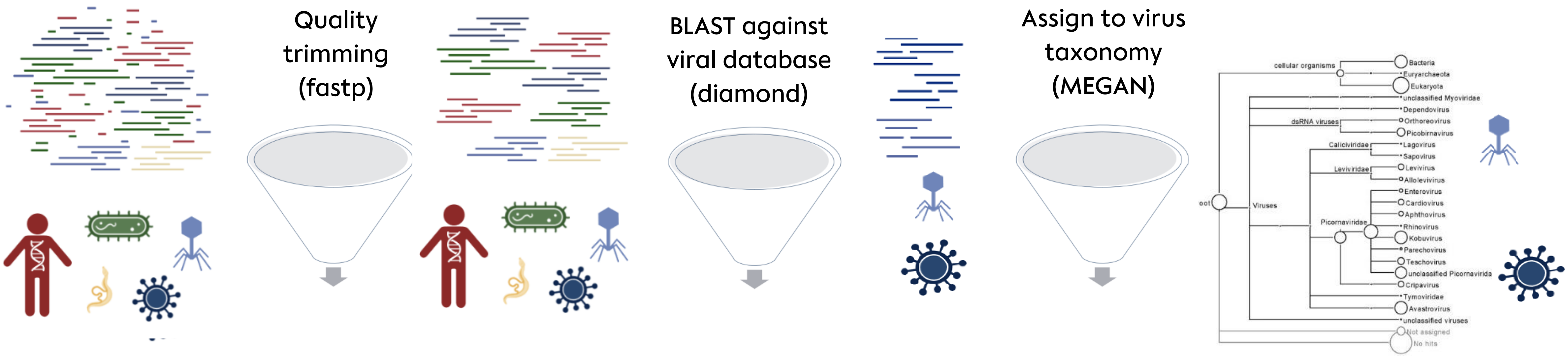
[1] Center for Bioinformatics, Tübingen University, Sand 14, 72076 Tübingen, Germany; [2] Center for Comparative Genomics and Bioinformatics, Center for Infectious Disease Dynamics, Penn State University, University Park, Pennsylvania 16802, USA

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.]

Meanwhile...

Introduction of the tools used in the pipeline



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

- 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

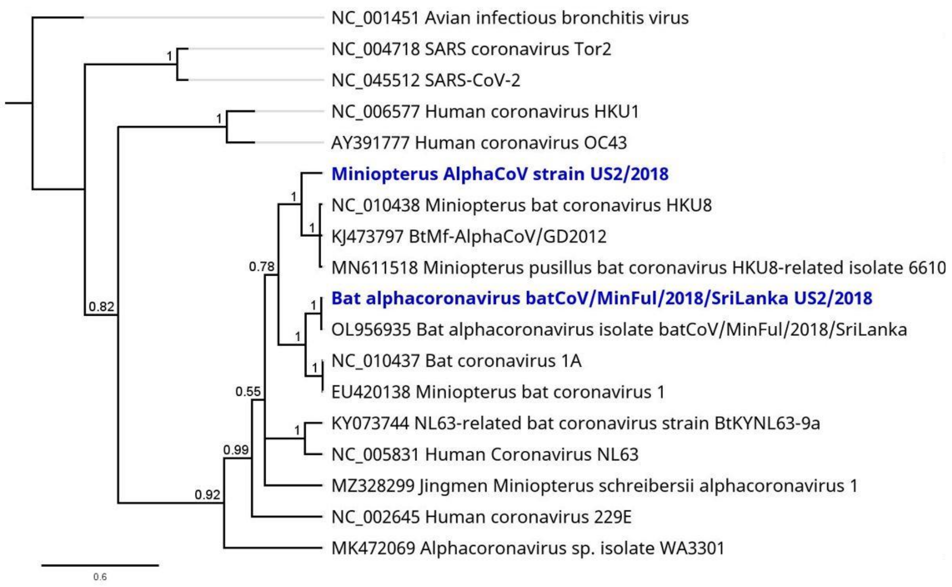

NC_001451 Avian infectious bronchitis virus
NC_004718 SARS coronavirus Tor2
NC_045512 SARS-CoV-2
NC_006577 Human coronavirus HKU1
AY391777 Human coronavirus OC43
Miniopterus AlphaCoV strain US2/2018
NC_010438 Miniopterus bat coronavirus HKU8
KJ473797 BtMf-AlphaCoV/GD2012
MN611518 Miniopterus pusillus 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
0.78
0.82
0.55
0.99
0.92
0.6

Figure 1: Example phylogenetic tree

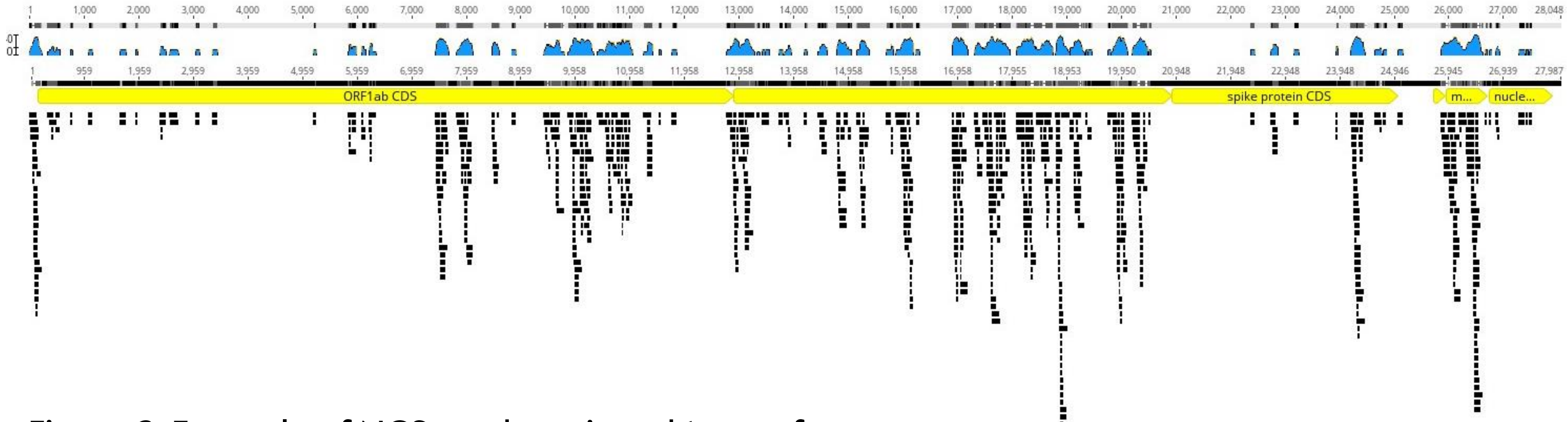


Figure 2: Example of NGS reads assigned to a reference genome