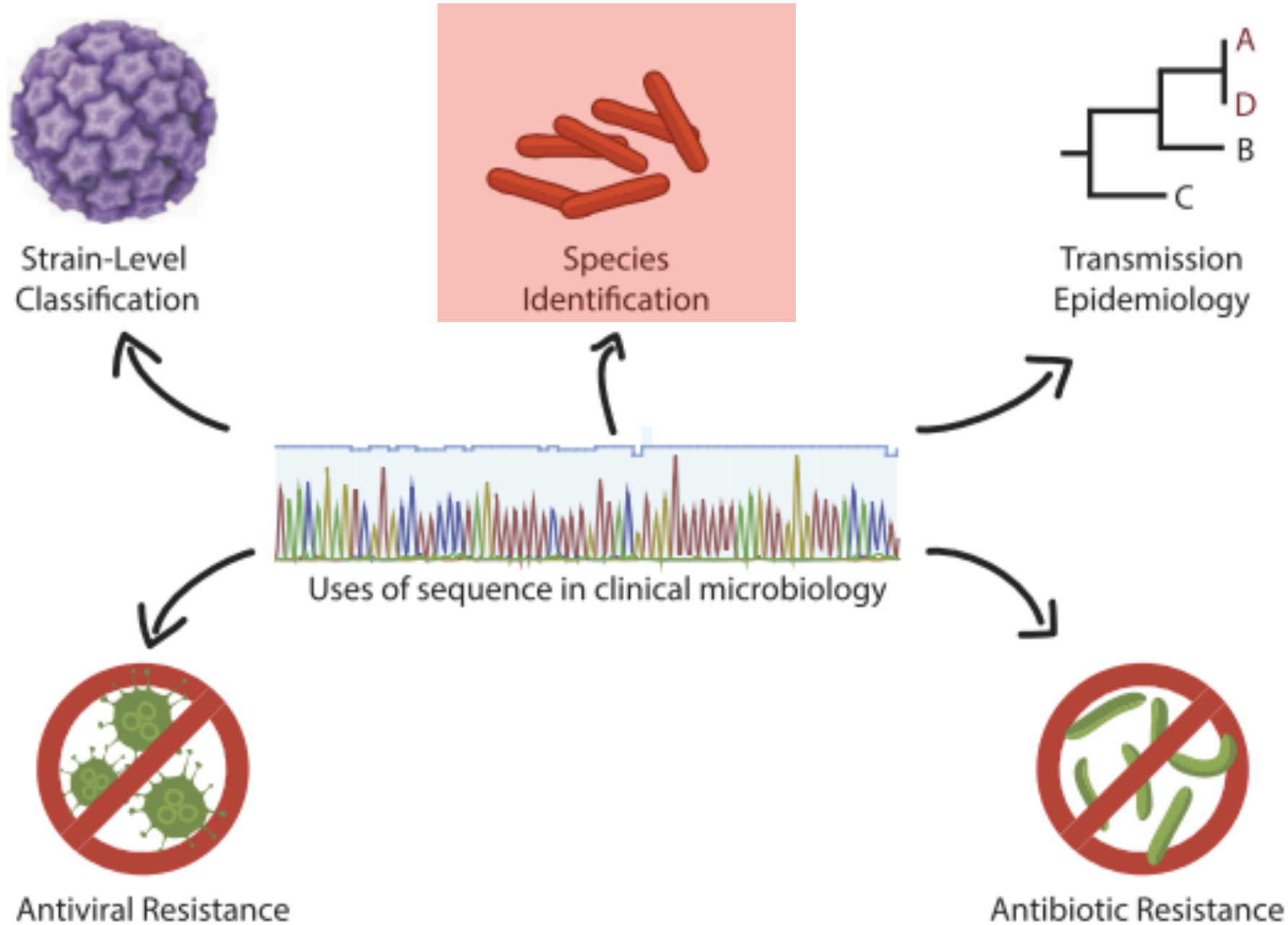
Judy Breuer's lab
Institute of Child Health

Introduction to viral metagenomics

Dr. Cristina Venturini
c.venturini@ucl.ac.uk
@cristina_ventu

Why?



The challenge of diagnostic metagenomics
Greninger et al, 2018. Expert Review of Molecular Diagnostics.

What is metagenomics?

- Metagenomics is a field of NGS that enables identification of microbial communities, and genetic detection, identification, and characterisation of disease-causing agents

Introduction

- Metagenomics approaches have been used extensively for pathogen discovery and for the characterisation of microbial diversity in environmental and clinical samples

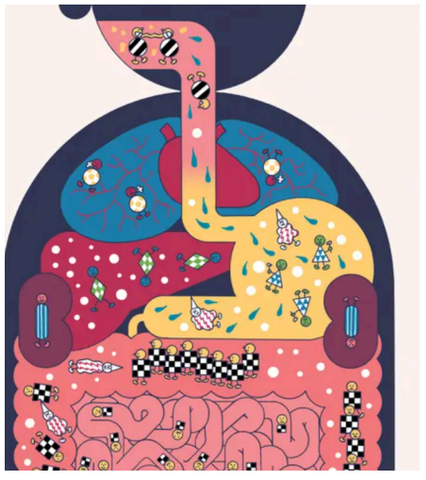
- Total DNA and/or RNA are extracted and sequence from a sample and compare to reference genome database to identify pathogens (viruses, bacteria, fungi)

Applications of metagenomics

Environmental

Medical

Microbiome studies

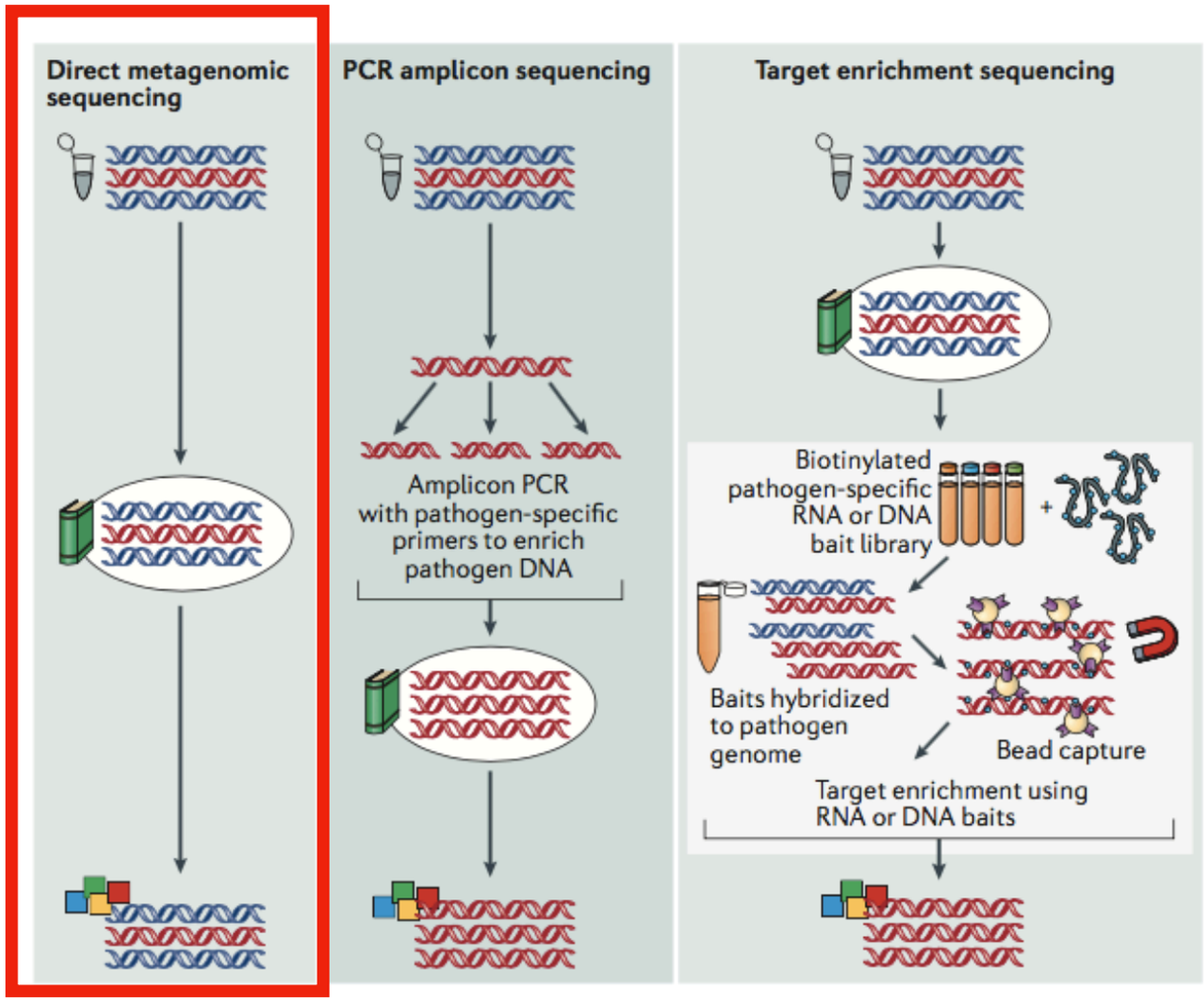


Surveillance & discovery

Diagnostic

Images from the Noun project + The Guardian article

In the lab…



Clinical and biological insights from viral genome sequencing
Houldcroft CJ, 2017
Nature Reviews Microbiology

- All specimens originally include a mix of host (in blue) and pathogen (in red) DNA sequences.

- Direct metagenomics sequencing provides an accurate representation of the sequences in the samples

- but: sequencing, data analysis and storage cost

In the lab…

- **Advantages:**
    - It doesn’t require any prior knowledge of the genomes under study for primers or probe design
    - It allows identification of all pathogens present in a wide variety of samples (i.e. cerebrospinal fluid, sputum, serum, stool, amniotic fluid..)
- **Disadvantages:**
    - Large quantities of genomic starting material (e.g. a high viral load)
        - Underrepresentation/loss of minority genomes (low sensitivity) as DNA from the host genome and commensal microorganisms are also amplified
        - Risk of contamination during sample collection and the analytical process

In the lab…16s vs NGS

- 16s: rRNA gene: section of prokaryotic DNA found in all bacteria.
  - Fast and cost effective
  - Community profiling/microbial ecology (operational taxonomic units OTUs)
- NGS: whole metagenome
  - Slower and more expensive
  - Additional analysis: species/strain identification, variant analysis, resistance
  - Can ID viruses

Data Analysis - pipeline

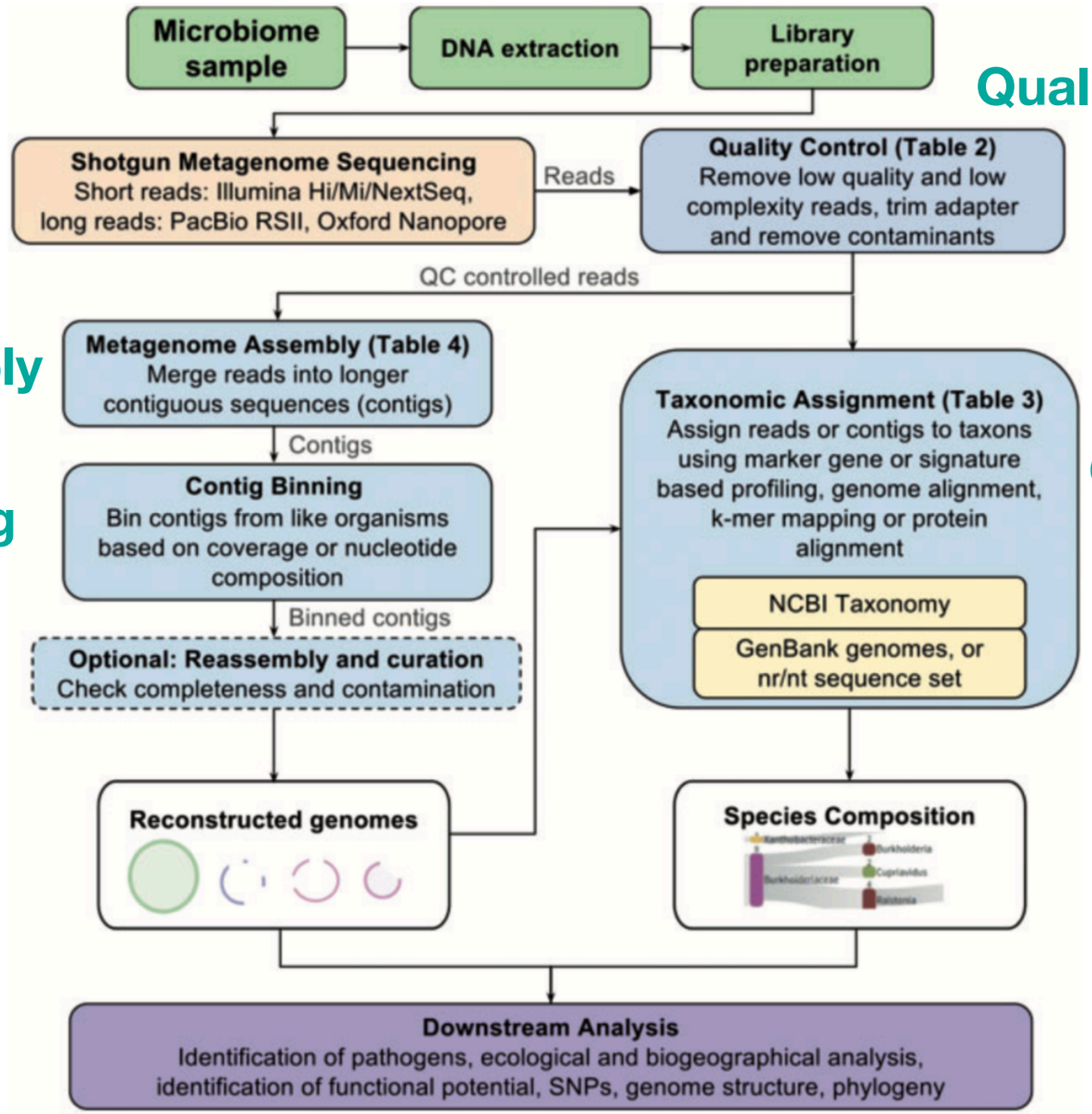
Quality control

Assembly

Optional

Binning

Classification



A review of methods and databases for metagenomic classification and assembly.
Breitwieser F, 2017.
Briefings in Bioinformatics

Quality control

- The first step is to perform **sequence QC** to remove technical errors from the analysis.

- Why? We need to remove adapter sequences, excessively short reads, low-quality reads

Table 1. Bioinformatic programs for data quality control in short-read and long-read sequencing. All websites/GitHub Repository links were accessed on 23 December 2022.

Program	Type of Data	Website/GitHub Repository
FastQC	Short reads	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
MultiQC	Short reads	https://multiqc.info
LongQC	Long reads	https://github.com/yfukasawa/LongQC
MinionQC	Long reads	https://github.com/roblanf/minion_qc

Table 2. Bioinformatic programs for data trimming in short-read and long-read sequencing. All websites/GitHub Repository links were accessed on 23 December 2022.

Program	Type of Data	Website/GitHub Repository
Trimmomatic	Short reads	http://www.usadellab.org/cms/?page=trimmomatic
Fastp	Short reads	https://github.com/OpenGene/fastp
Cutadapt	Short reads	https://cutadapt.readthedocs.io/en/stable/
SOAPnuke	Short reads	https://github.com/BGI-flexlab/SOAPnuke
NanoPack	Long reads	https://github.com/wdecoster/nanopack
SequelTools	Long reads	https://github.com/ISUgenomics/SequelTools

Quality control

- The second step is to eliminate reads of no interest i.e. host genomes and contaminants

- Why? To decrease computational time, reduce false positives and avoid assembly of chimeric virus-host sequences

- Strategies:

  - Read mappers can be applied to remove all sequences mapping a selected reference genome (i.e. BWA, Bowtie2, BBMap, Minimap 2) (https://github.com/Finn-Lab/Metagen-FastQC)

  - Other tools identify specific sequences belonging to specific taxa i.e. sequences are passed through a filtering only selecting reads with certain features (i.e. VirusHunter)

  - In some situations other RNA sequence types may need to be removed (i.e. ribosomal RNA)

  - We could also do a first round of taxonomic profiling of the reads before the assembly. We can then select sequences belonging to viruses and continue with further analysis (i.e. kraken2)

Assembly

Reads vs assembled contigs

- **Reads (1):**
    - quantitative community profiling
    - identification of organisms with close relatives in the db
    - clinical microbiology: presence/absence of infectious pathogen
- **assembled contigs (denovo assembly) (2)**:
    - no close relative of a specie
    - more qualitative understanding

Assembly

- To identify the virus present, we must first restore the metagenomes. We need to generate **contigs,** sets of sequences that can be overlapped to provide a longer, continuous sequence.

- Why do we use denovo assembly?

    - target genome is poorly characterised
    - unknown entities in metagenomes

- Challenge: coverage varies across both among different genomes and within individual genomes

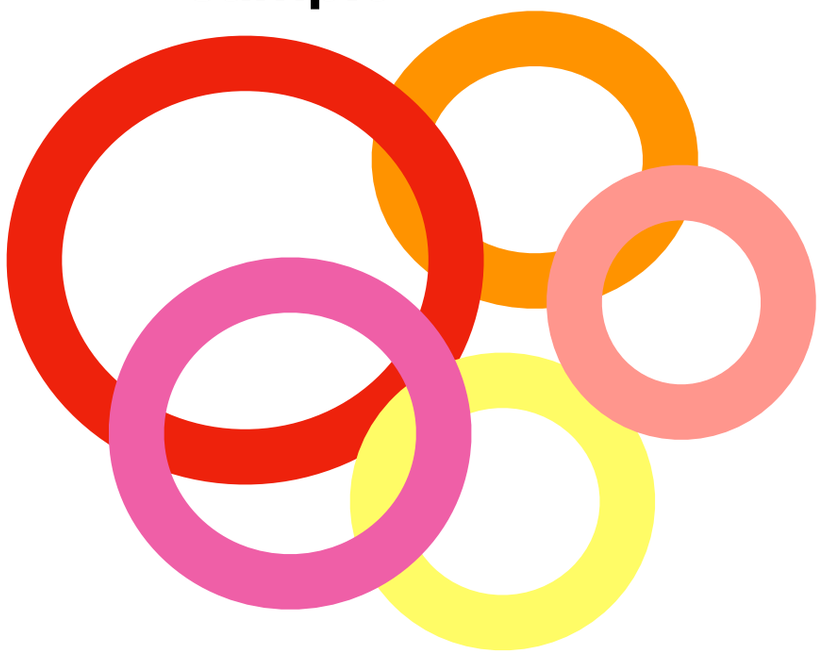
Table 4. Bioinformatic tools for metagenome assembly for short-read, long-read, and hybrid assemblies. All websites/GitHub Repository links were accessed on 23 December 2022.

Program	Read Length	Algorithm	Website/GitHub Repository
MEGAHIT	Short reads	De Bruijn graph	https://github.com/voutcn/megahit
metaSPADES	Short reads	De Bruijn graph	https://github.com/ablab/spades
IDBA-UD	Short reads	De Bruijn graph	https://i.cs.hku.hk/~alse/hkubrg/projects/idba_ud/
MetaVelvet	Short reads	De Bruijn graph	http://metavelvet.dna.bio.keio.ac.jp
Omega2	Short reads	Overlap layout consensus	https://github.com/qiumingyao/omega2
metaFlye	Long reads	Overlap layout consensus	https://github.com/fenderglass/Flye
Canu	Long reads	Overlap layout consensus	https://github.com/marbl/canu
NECAT	Long reads (Nanopore)	String graph	https://github.com/xiaochuanle/NECAT
HybridSPADES	Hybrid	De Bruijn graph	https://github.com/ablab/spades
OPERA-MS	Hybrid	De Bruijn graph	https://github.com/CSB5/OPERA-MS
HASLR	Hybrid	De Bruijn graph	https://github.com/vpc-ccg/haslr
Wegan	Hybrid	Synthetic scaffolding graph	https://github.com/adigenova/wengan

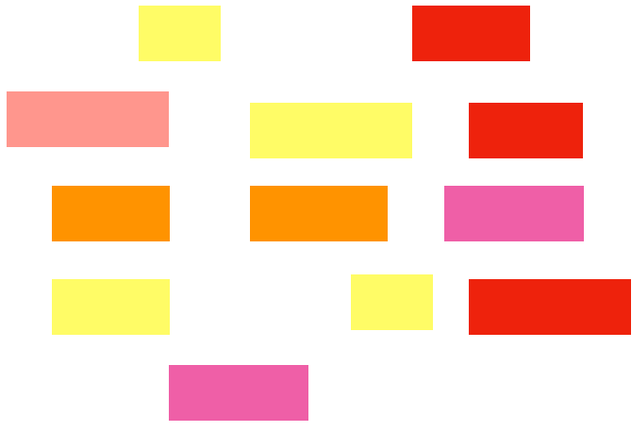
https://doi.org/10.3390/v15020587

Assembly

Viral genomes present in a sample



Viral reads



Don’t forget to check the quality of your assemblies!

Viral contigs

Classification

Sequence classification: taxonomic profilers match sequences against a database of microbial genomes to identify the taxon of each sequence

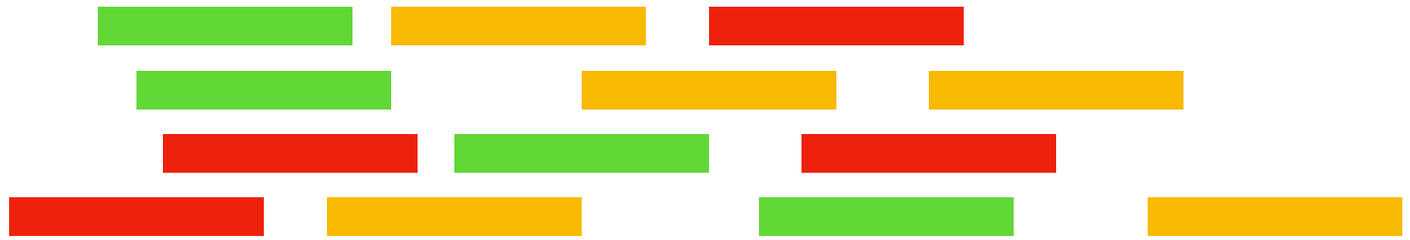
- Alignment-based: BLAST (e.g. searching similarity in sequences)
- K-mer approaches: Kraken, Centrifuge, Kallisto
- Protein-based programs: they translate sequences to enable comparisons with reference protein db (i.e. DIAMOND)
- Bayesian mixture models: Metamix
- Neural networks algorithms (i.e. DeepVirFinder)

https://doi.org/10.3390/v15020587

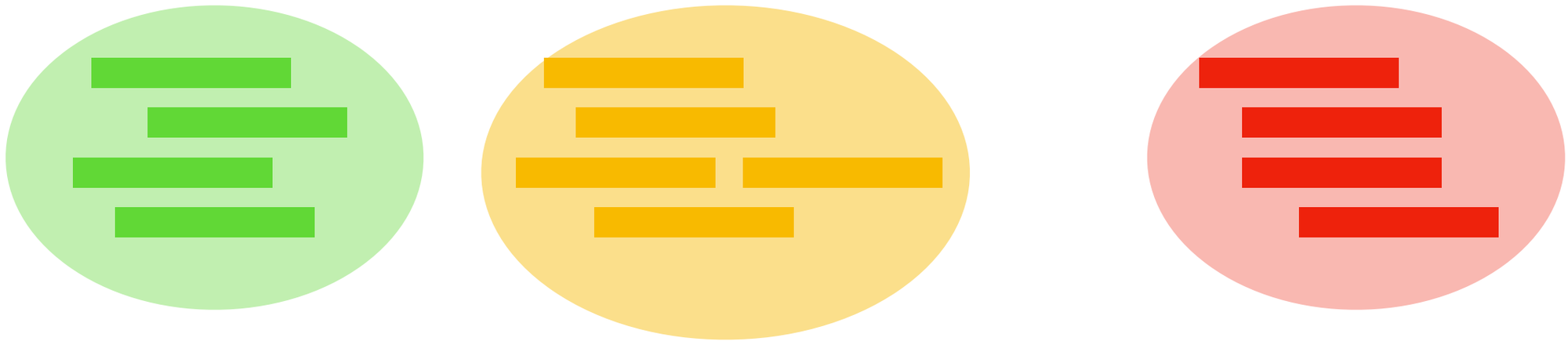
Classification

Fungal
Bacterial
Viral

Reads/contigs



Binning



Fungal reads/contigs

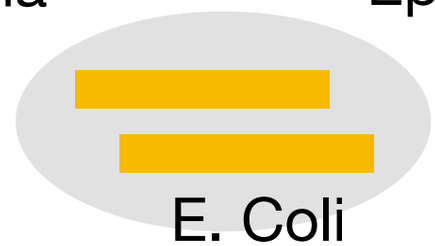
Bacterial reads/contigs

Viral reads/contigs

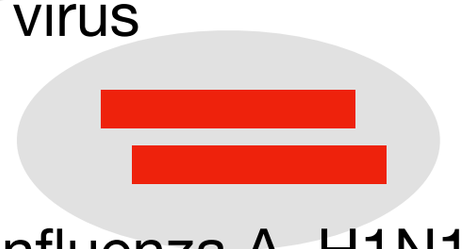
Classification

S. Pneumonia

Epstein-Barr virus



E. Coli



Influenza A, H1N1

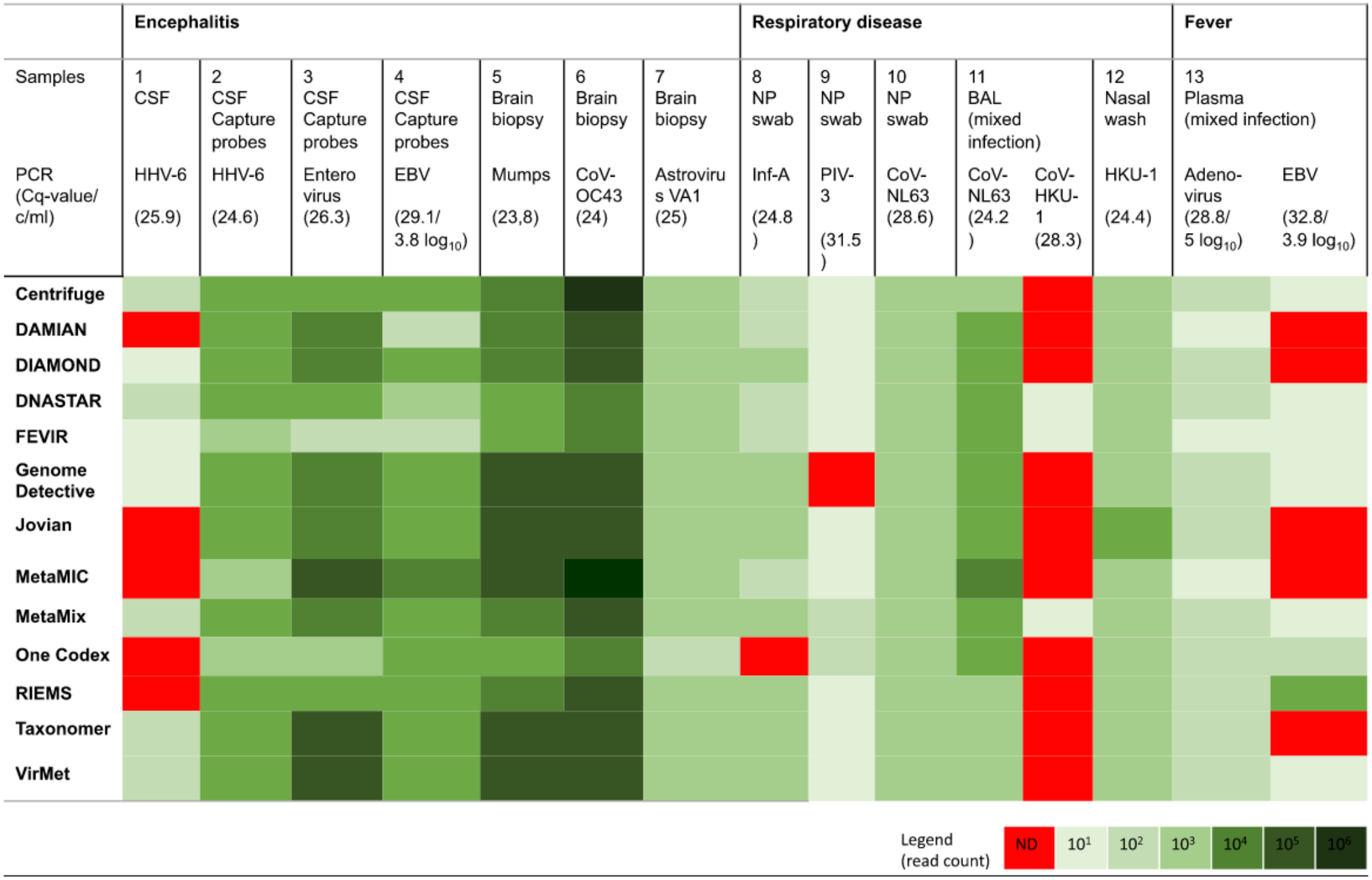
Classification

Table 6. Bioinformatic taxonomic profilers, including algorithm, use of custom databases, and website or GitHub repository. All websites/GitHub Repository links were accessed on 23 December 2022.

Program	Type of Classifier	Allows Custom Databases	Website/GitHub Repository
Kraken2	k-mers	Yes	https://ccb.jhu.edu/software/kraken2/
Kraken-HLL	k-mers	Yes	https://github.com/Krischan/krakenhll
KrakenUniq	k-mers	Yes	https://github.com/fbreitwieser/krakenuniq
Centrifuge	k-mers	Yes	https://github.com/DaehwanKimLab/centrifuge
Ganon	k-mers	Yes	https://github.com/pirovc/ganon
Bracken	k-mers	Yes	https://github.com/jenniferlu717/Bracken
MetaCache	k-mers	Yes	https://github.com/muellan/metacache
CLARK	k-mers	Yes	http://clark.cs.ucr.edu
VirusTaxo	k-mers	No	https://omics-lab.com/virustaxo
Metavir2	k-mers	No	https://github.com/jhayer/metavir
k-SLAM	k-mers	Yes	https://github.com/aindj/k-SLAM
Taxonomer	k-mers	No	http://taxonomer.com
LMAT	k-mers	No	https://computing.llnl.gov/projects/livermore-metagenomics-analysis-toolkit
Sourmash	k-mers	No	https://sourmash.readthedocs.io/en/latest/
metaOthello	k-mers	Yes	https://github.com/xa6xa6/metaOthello
ProPhyle	k-mers	No	https://prophyle.github.io
TaxMaps	k-mers	Yes	https://github.com/nygenome/taxmaps#sge
Kaiju	Protein-coding	Yes	https://kaiju.binf.ku.dk
DIAMOND	Protein-coding	Yes	https://github.com/bbuchfink/diamond
MMseqs2	Protein-coding	Yes	https://github.com/soedinglab/MMseqs2
IGGsearch	Marker gene	No	https://github.com/snayfach/IGGsearch
MetaPhlAn3	Marker gene	No	https://huttenhower.sph.harvard.edu/metaphlan
GOTTCHA	Marker gene	No	https://lanl-bioinformatics.github.io/GOTTCHA/
DeepVirFinder	CNN	No	https://github.com/jessieren/DeepVirFinder
BLAST	Alignment-based	Yes	https://blast.ncbi.nlm.nih.gov/Blast.cgi
DUDes	DUD	No	https://github.com/pirovc/dudes
MCP	Alignment-based	Yes	https://microba.com/microbiome-research/

https://doi.org/10.3390/v15020587

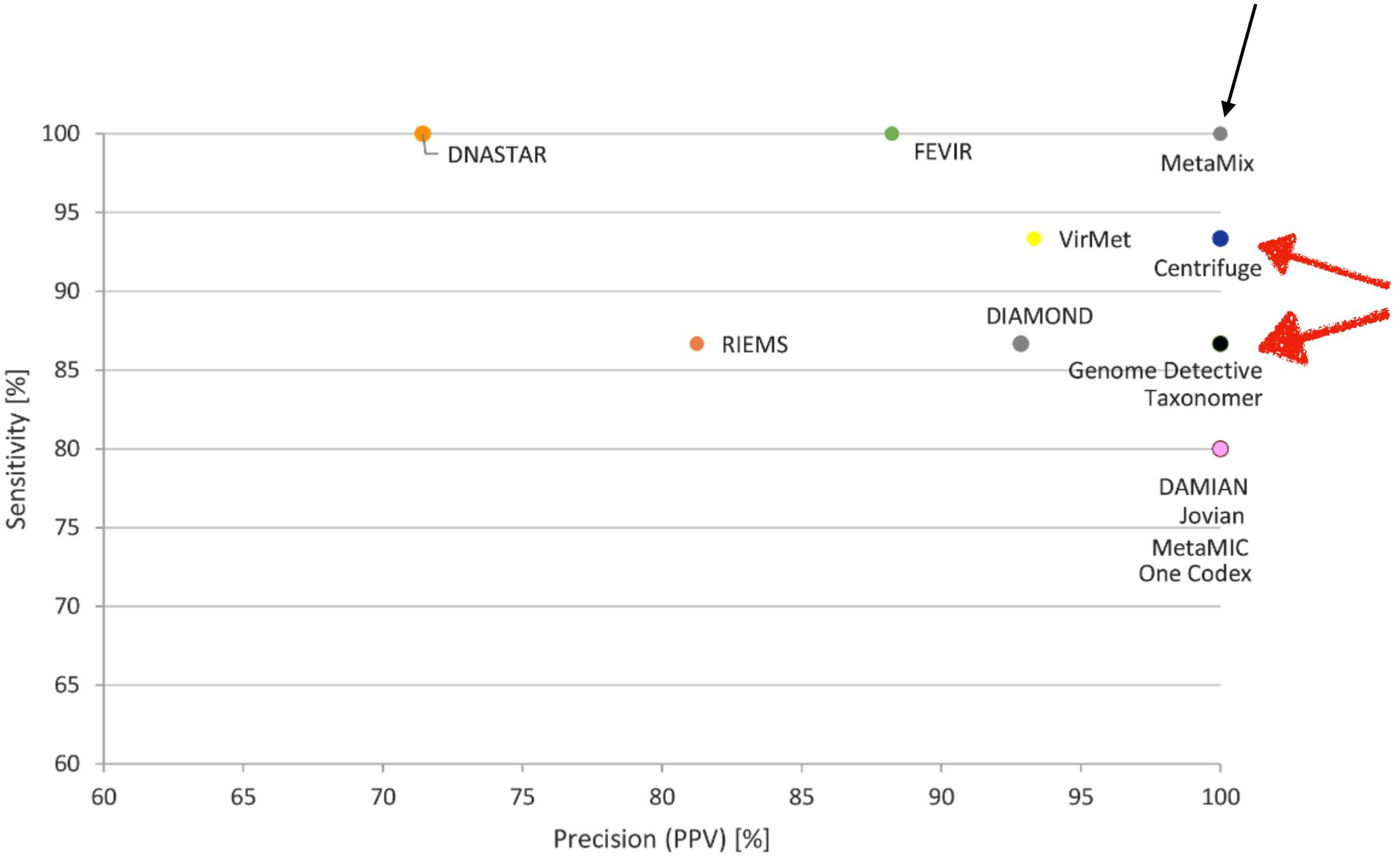
Choosing the classification method



Benchmark of thirteen bioinformatic pipeline for metagenomics diagnostics using datasets from clinical samples

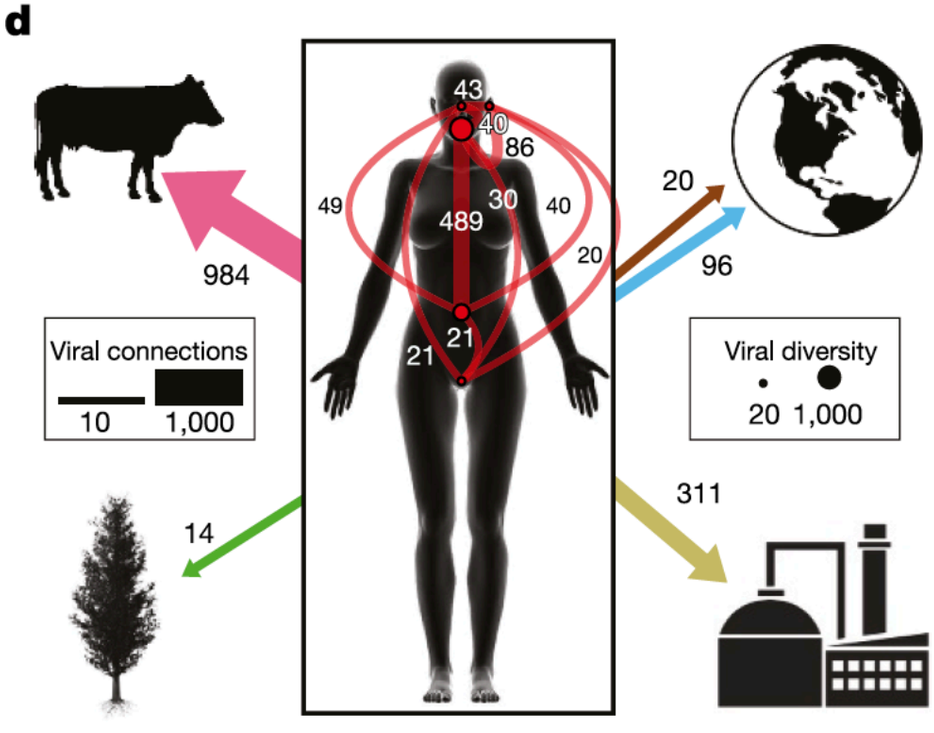
JJC de Vries et al., 2021

Journal of Clinical Virology


100
95
90
85
80
75
70
65
60
Sensitivity [%]
60
65
70
75
80
85
90
95
100
Precision (PPV) [%]
DNASTAR
FEVIR
MetaMix
VirMet
Centrifuge
DIAMOND
RIEMS
Genome Detective
Taxonomer
DAMIAN
Jovian
MetaMIC
One Codex

Genome resources

- Database of genomes and on taxonomy of species
- NCBI Taxonomy database (GenBank, EMBL and DDBJ)
- Viral diversity problem
- > 125000 new DNA viruses (Paez-Espino,2016, Nature)
- International Committee on Taxonomy of Viruses



body-site specificity

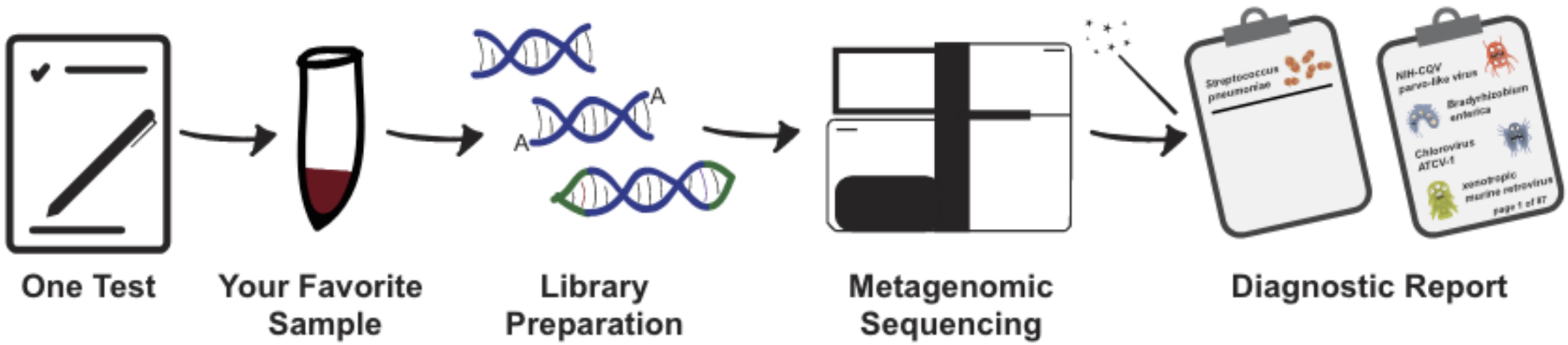
Genome resources

- **GenBank**
  - Relies on correct taxonomic identification and annotation provided by the submitter
  - Incorrect species name
  - “draft” genomes (contaminant contigs)
- **Refseq**
  - More curated genome resource
  - > 7000 viral genomes (May 2017)

Uses of metagenomics:

- Clinical diagnostics
- Pathogen discovery
- Microbiome and virome
- Drug resistance
- Virulence

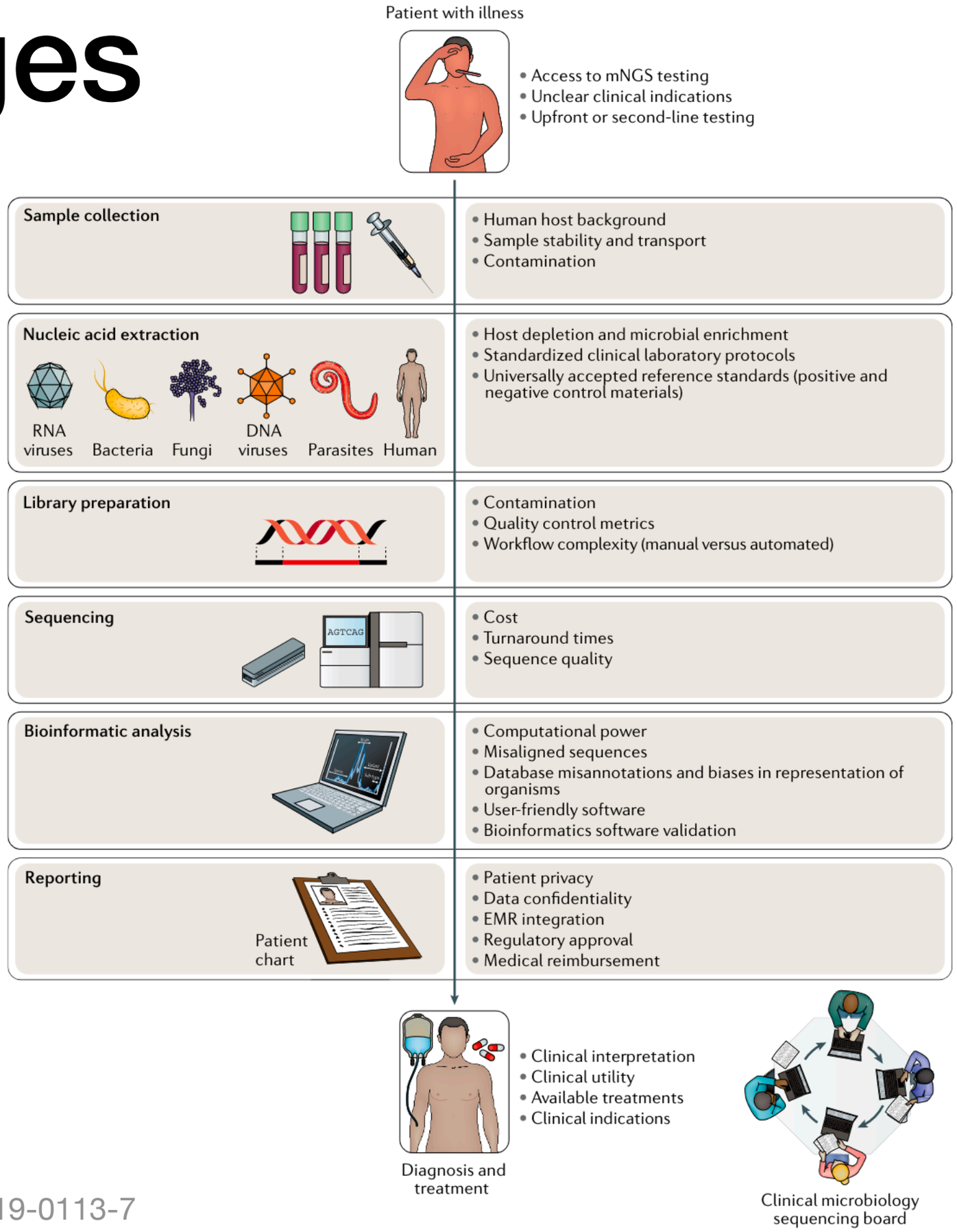
Clinical diagnostics (1)



The dream!

The challenge of diagnostic metagenomics
Greninger et al, 2018. Expert Review of Molecular Diagnostics.

Challenges



https://doi.org/10.1038/ s41576-019-0113-7

Clinical diagnostics (1)

- For several infectious diseases: aetiology is not known
- Inadeguate current diagnostic tests (PCR): rapid & sensitive BUT require prior knowledge
- Lack of timely diagnosis: poor patient management/ prognosis

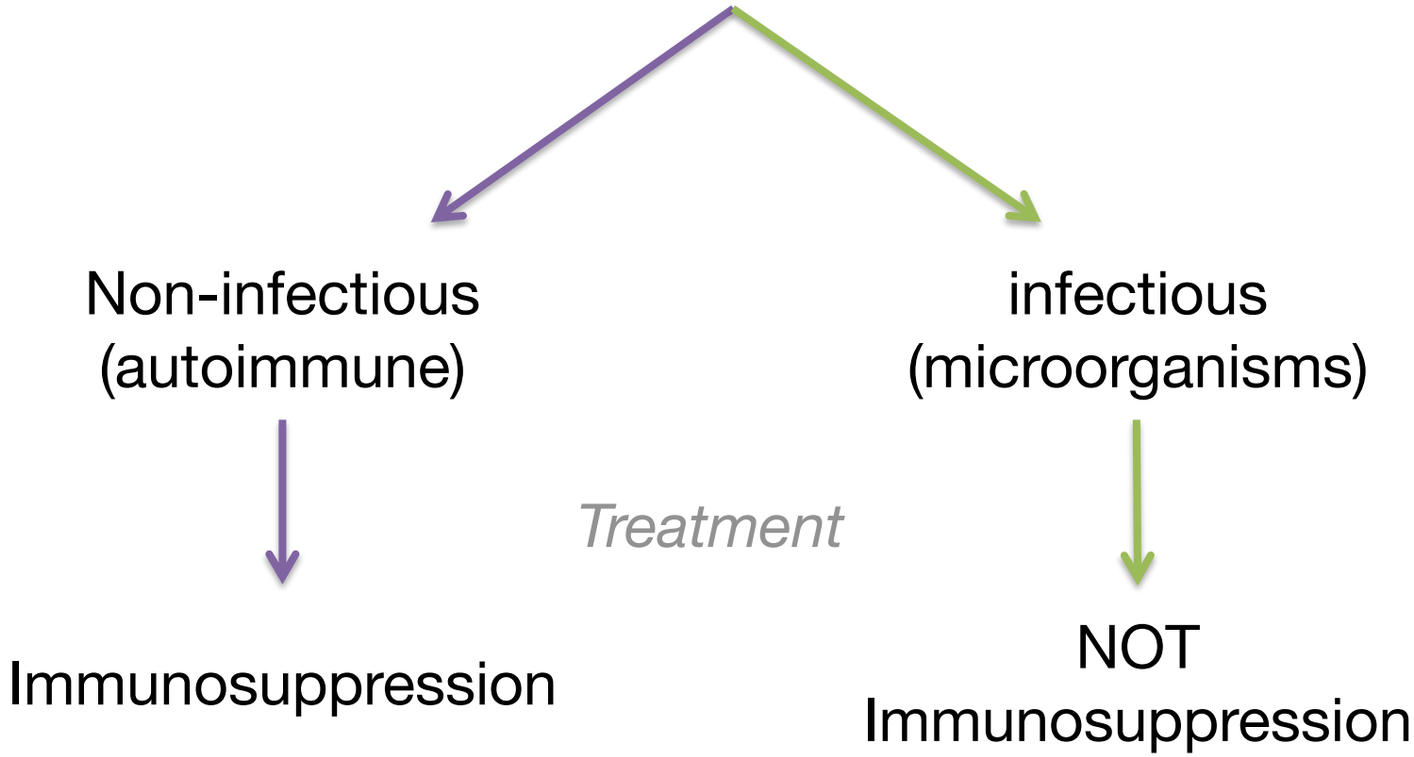
Prolonged paediatric fever 20%

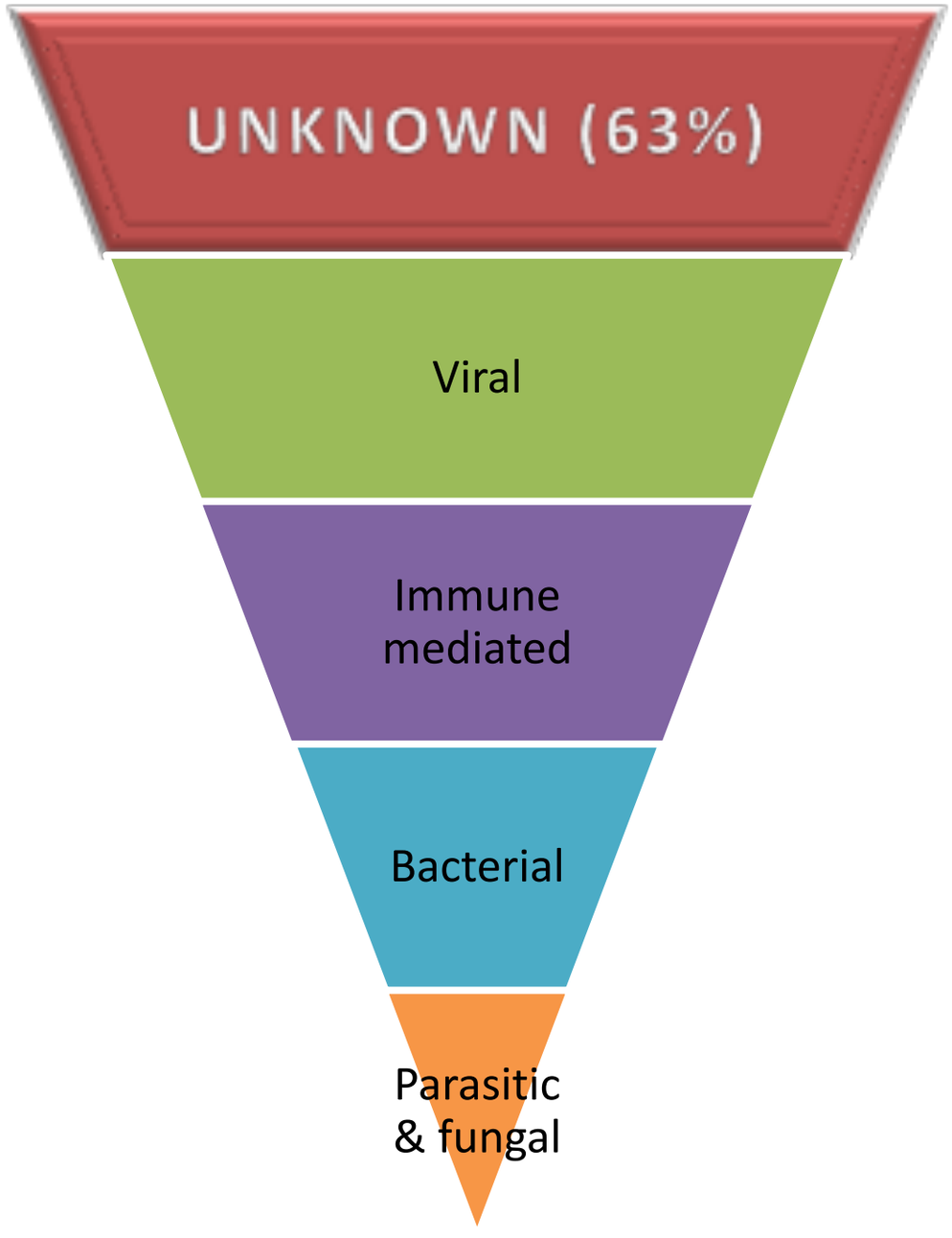
encephalitis 50-60%

gastroenteritis 50%

Encephalities

Up to 63% of encephalitis cases
no causative agent identified





Cases 1&2

Clinical Infectious Diseases

Issues More Content Publish Purchase Advertise About All Clinical Infectious Di

Volume 60, Issue 6
15 March 2015

Article Contents

Astrovirus VA1/HMO-C: An Increasingly Recognized Neurotropic Pathogen in Immunocompromised Patients

Julianne R. Brown ✉, Sofia Morfopoulou, Jonathan Hubb, Warren A. Emmett, Winnie Ip, Divya Shah, Tony Brooks, Simon M. L. Paine, Glenn Anderson, Alex Virasami, ... Show more

Clinical Infectious Diseases, Volume 60, Issue 6, 15 March 2015, Pages 881–888,
https://doi.org/10.1093/cid/ciu940
Published: 07 January 2015 **Article history**

- Astrovirus HAstV-VA1/HMO-C-UK1(a)
- Highly divergent from human astrovirus (HAstV 1–8))
- Closely related to VA1/HMO-C astroviruses, including one recovered from a case of fatal encephalitis in an immunosuppressed child.

Human coronavirus (OC43 (HCoV-OC43) – first report of causing encephalitis in humans

The NEW ENGLAND JOURNAL of MEDICINE

HOME ARTICLES & MULTIMEDIA ISSUES SPECIALTIES & TOPICS FOR AUTHORS CM

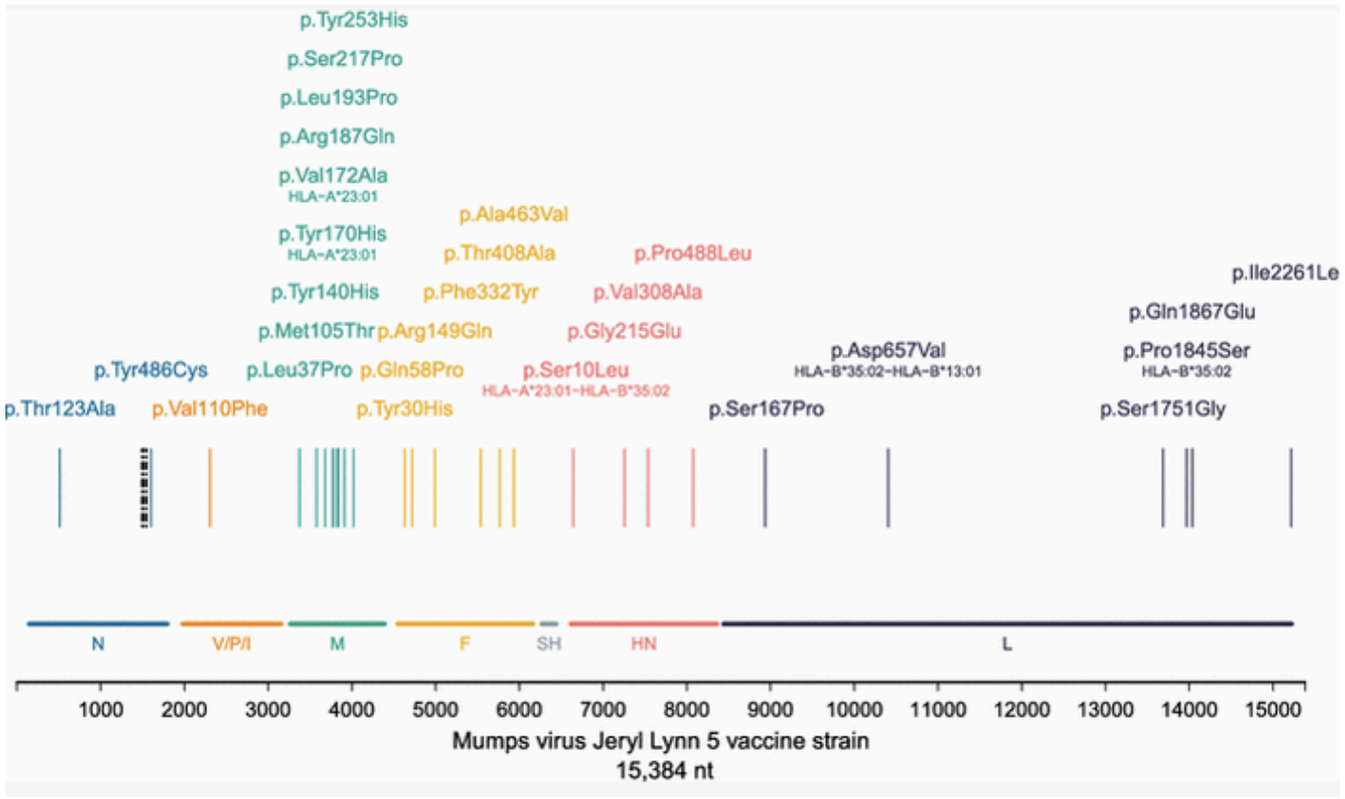
CORRESPONDENCE

Human Coronavirus OC43 Associated with Fatal Encephalitis

N Engl J Med 2016; 375:497-498 | August 4, 2016 | DOI: 10.1056/NEJMc1509458

Case 3

- 18 month old boy, SCID
- stem cell transplant
- 40 months post transplant, hospitalisation as acutely unwell – extensive screening and brain biopsy
- NextSeq: 100 million 2x81 base-paired run
- Metamix summary: 77,624 mumps reads
- Confirmed with PCR and IHC

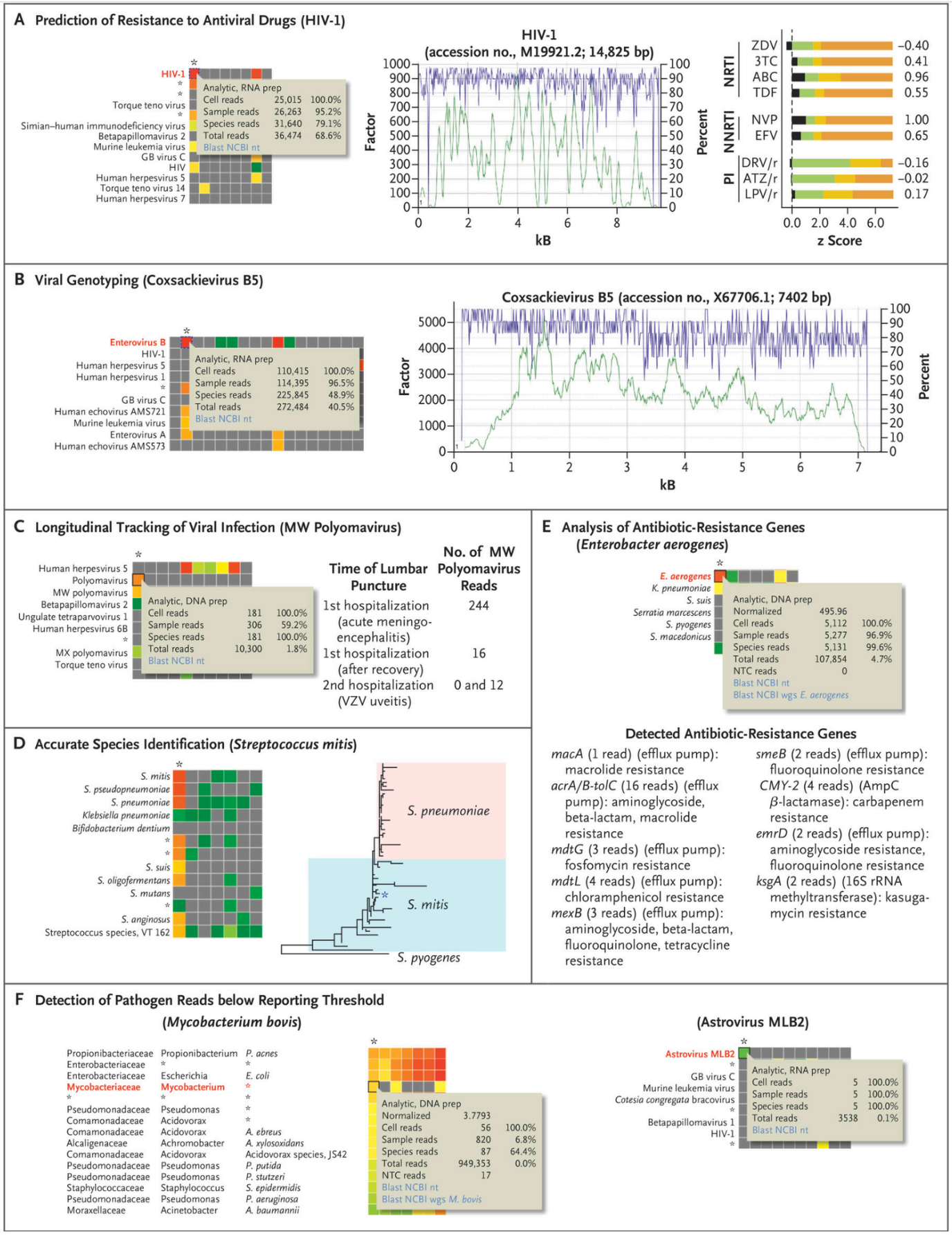


- The full-length viral sequence was recovered
- Viral sequence clustered closely with the MuV^{JL5} vaccine strains
- 28 missense amino acid substitutions between MuV-London and the vaccine
- T to C biased hypermutations in M gene– similar cases in of measles MMPE
- Strong case for deep sequencing of brain tissue

Clinical Metagenomic Sequencing for Diagnosis of Meningitis and Encephalitis

Wilson M, 2019.

New England Journal of Medicine



Pathogen discovery/surveillance (2)

CCDC weekly

A Novel Coronavirus Genome Identified in a Cluster of Pneumonia Cases — Wuhan, China 2019–2020

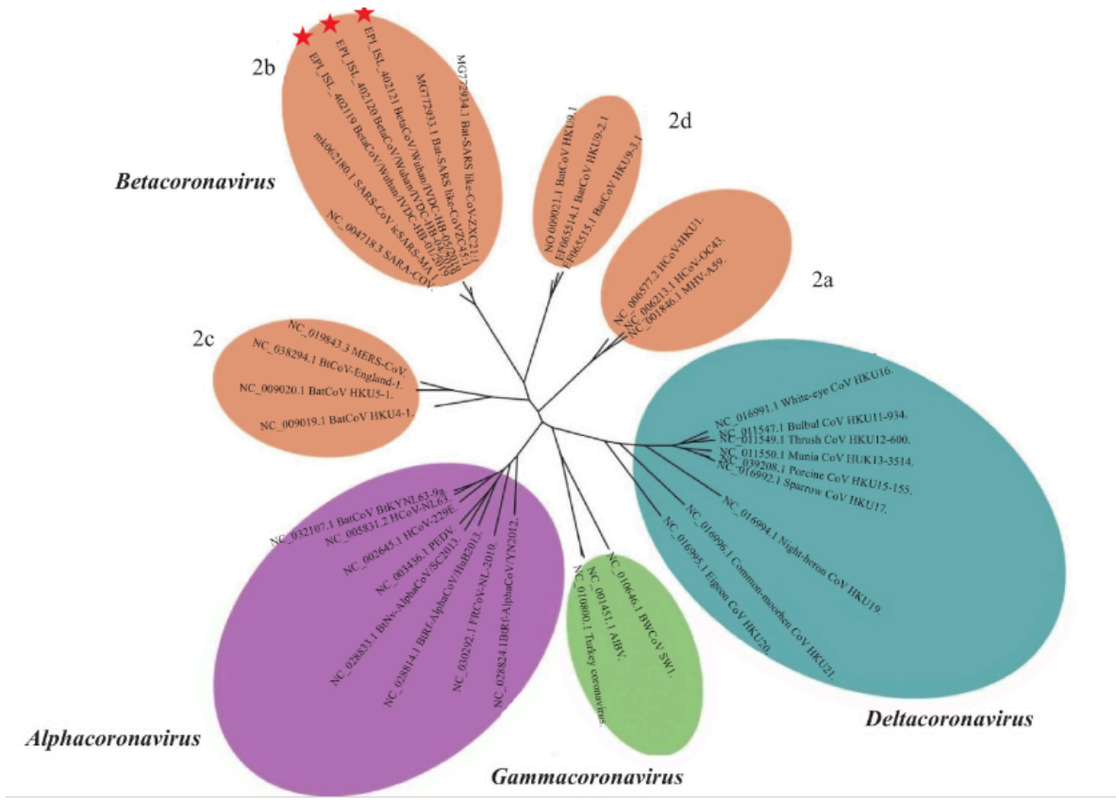
Wenjie Tan[1,2]; Xiang Zhao[1]; Xuejun Ma[1]; Wenling Wang[1]; Peihua Niu[1]; Wenbo Xu[1]; George F. Gao[1]; Guizhen Wu[1,2]

Article | Open Access | Published: 03 February 2020

A pneumonia outbreak associated with a new coronavirus of probable bat origin

Peng Zhou, Xing-Lou Yang, ... Zheng-Li Shi + Show authors

Nature **579**, 270–273 (2020) | Cite this article



Article | Open Access | Published: 03 February 2020

A new coronavirus associated with human respiratory disease in China

Fan Wu, Su Zhao, Bin Yu, Yan-Mei Chen, Wen Wang, Zhi-Gang Song, Yi Hu, Zhao-Wu Tao, Jun-Hua Tian, Yuan-Yuan Pei, Ming-Li Yuan, Yu-Ling Zhang, Fa-Hui Dai, Yi Liu, Qi-Min Wang, Jiao-Jiao Zheng, Lin Xu, Edward C. Holmes & Yong-Zhen Zhang

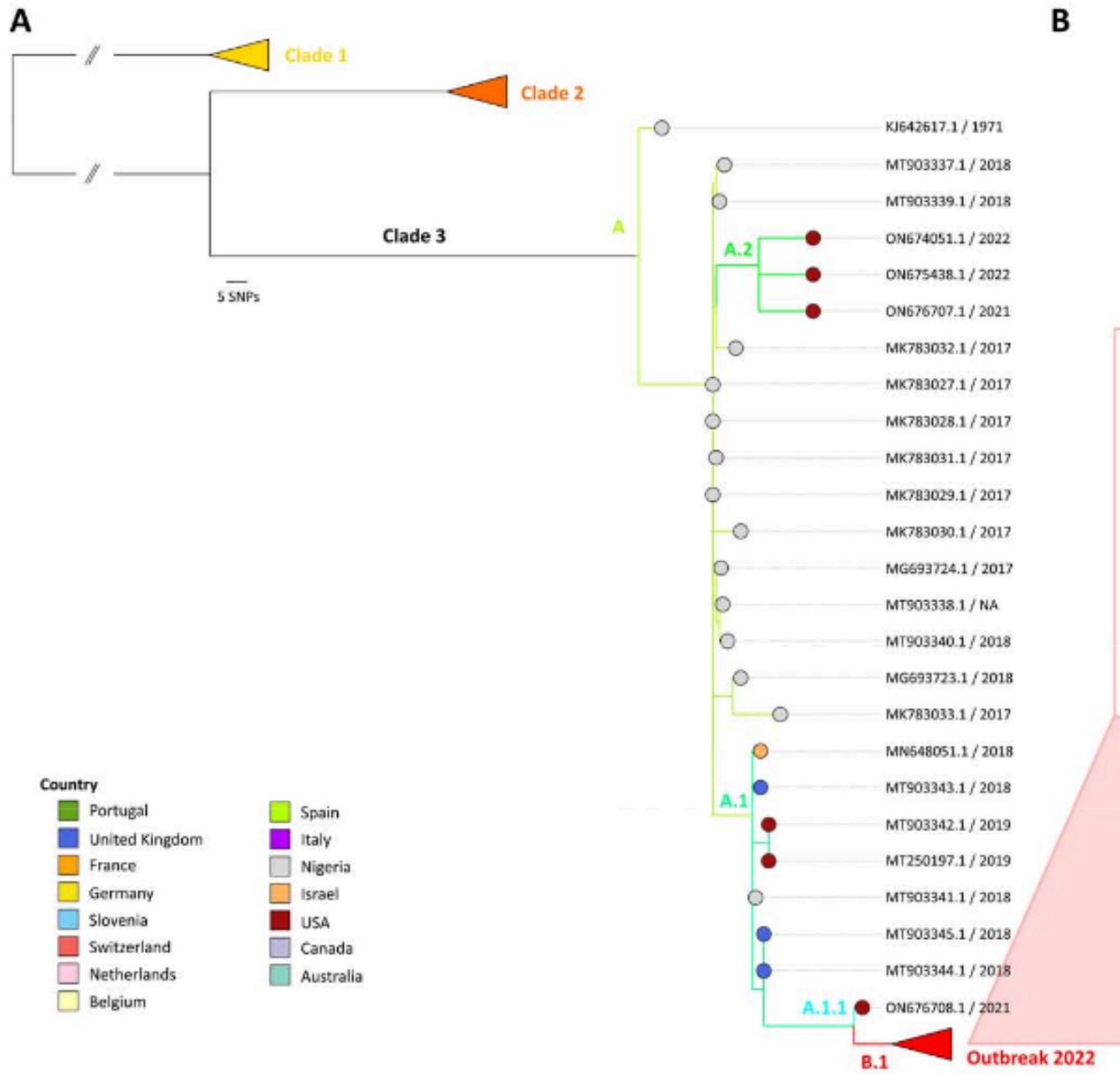
Nature **579**, 265–269 (2020) | Cite this article

Brief Communication | Published: 24 June 2022

Phylogenomic characterization and signs of microevolution in the 2022 multi-country outbreak of monkeypox virus

Joana Isidro, Vítor Borges, Miguel Pinto, Daniel Sobral, João Dourado Santos, Alexandra Nunes, Verónica Mixão, Rita Ferreira, Daniela Santos, Silvia Duarte, Luís Vieira, Maria José Borrego, Sofia Núncio, Isabel Lopes de Carvalho, Ana Pelerito, Rita Cordeiro & João Paulo Gomes

Nature Medicine (2022) | Cite this article



- Metagenomics allowed the rapid reconstructions and phylogenomic characterisation of the first MPXV outbreak genome sequences

- Single origin and clustered with 2018-2019 cases linked to an endemic country; however, it segregates in a divergent phylogenetic branch, likely reflecting continuous accelerated evolution

Metagenomics data analysis is challenging

- What species are there? How many? How much?

1. Sequencing technology

    - Short fragments from several microbes

2. Clinical sample

    - Low level of pathogen signal
    - Low biomass samples

3. Contamination in every step

    - Negative/positive controls

4. Reference databases

    - Incompleteness/errors

Slide by Sofia Morfopoulou

Tutorial

We'll be analysing a clinical specimen from a 41-years-old patient with no history of hepatitis, tubercolosis or diabetes. The patient reported fever, chest tightness, cough, pain and weakness and was admitted to hospital 6 days after the onset of the disease. Preliminary aetiological investigations excluded the presence of influenza virus, *Chlamydia pneumoniae* and *Mycoplasma pneumoniae* and this was confirmed by PCR. Other common respiratory pathogens, including human adenoviruses, also tested negative. To investigate the possible aetiological agents associated with this disease, bronchoalveolar lavage fluid (BALF) was collected and deep meta-transcriptomic was performed.

You need to analyse the data to determine the cause of the disease and determine what treatment options may be available.

```
cd
rm -r  Metagenomics_Training
```