bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io


Creative Commons
This page is available in the following languages:
Afrikaans български Català Dansk Deutsch Ελληνικά English English (CA) English (GB) English (US) Esperanto
Castellano Castellano (AR) Español (CL) Castellano (CO) Español (Ecuador) Castellano (MX) Castellano (PE)
Euskara Suomeksi français français (CA) Galego עברית hrvatski Magyar Italiano 日本語 한국어 Macedonian Melayu
Nederlands Norsk Sesotho sa Leboa polski Português română slovenski jezik српски srpski (latinica) Sotho svenska
中文 華語 (台灣) isiZulu
creative commons
Attribution-Share Alike 2.5 Canada
You are free:
to Share — to copy, distribute and transmit the work
to Remix — to adapt the work
Free Cultural
APPROVED FOR
Works
Under the following conditions:
Attribution. You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
Share Alike. If you alter, transform, or build upon this work, you may distribute the resulting work only under the same or similar licence to this one.
For any reuse or distribution, you must make clear to others the licence terms of this work.
Any of the above conditions can be waived if you get permission from the copyright holder.
The author's moral rights are retained in this licence.
Disclaimer
Your fair dealing and other rights are in no way affected by the above.
This is a human-readable summary of the Legal Code (the full licence) available in the following languages:
English French
Learn how to distribute your work using this licence

Emerging Pathogen Detection and Identification

Aaron Petkau
Infectious Disease Genomic Epidemiology
April 18-21, 2023

bioinformatics.ca

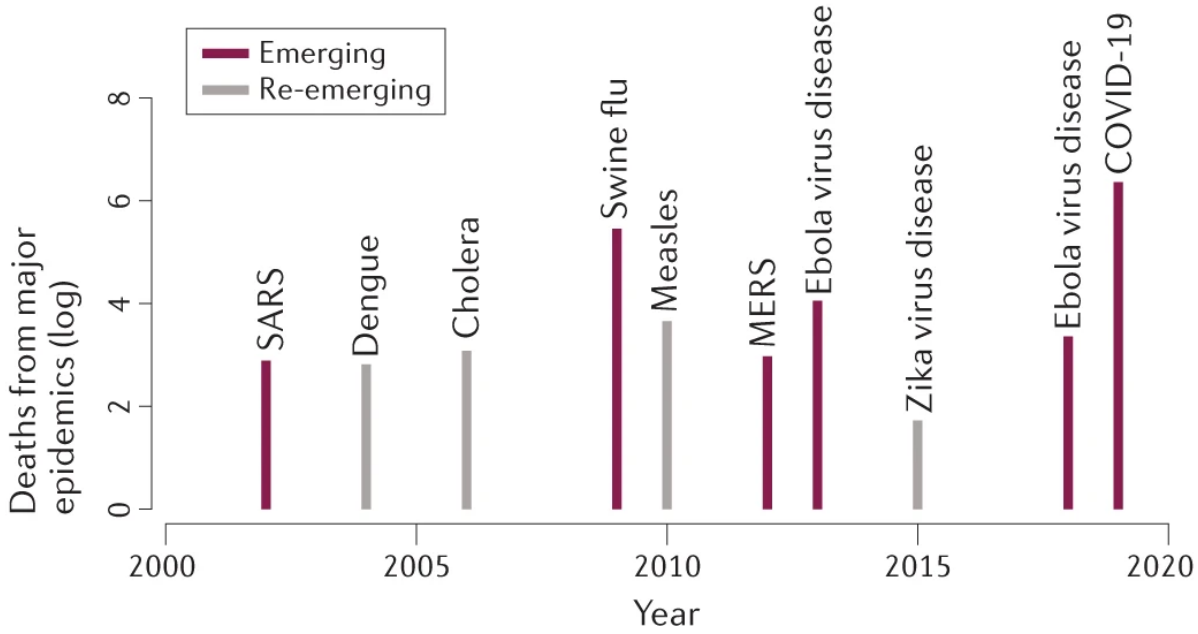
Public Health Agency of Canada
Agence de la santé publique du Canada

Learning Objectives

- By the end of this lecture, you will:
  - Learn the processes and techniques used to detect existing infectious diseases
  - Understanding metagenomics sequencing and the use of Kraken/Kraken2 for data analysis
  - Be able to identify a new pathogen using metagenomic sequence data

Novel and Emerging Infectious Diseases

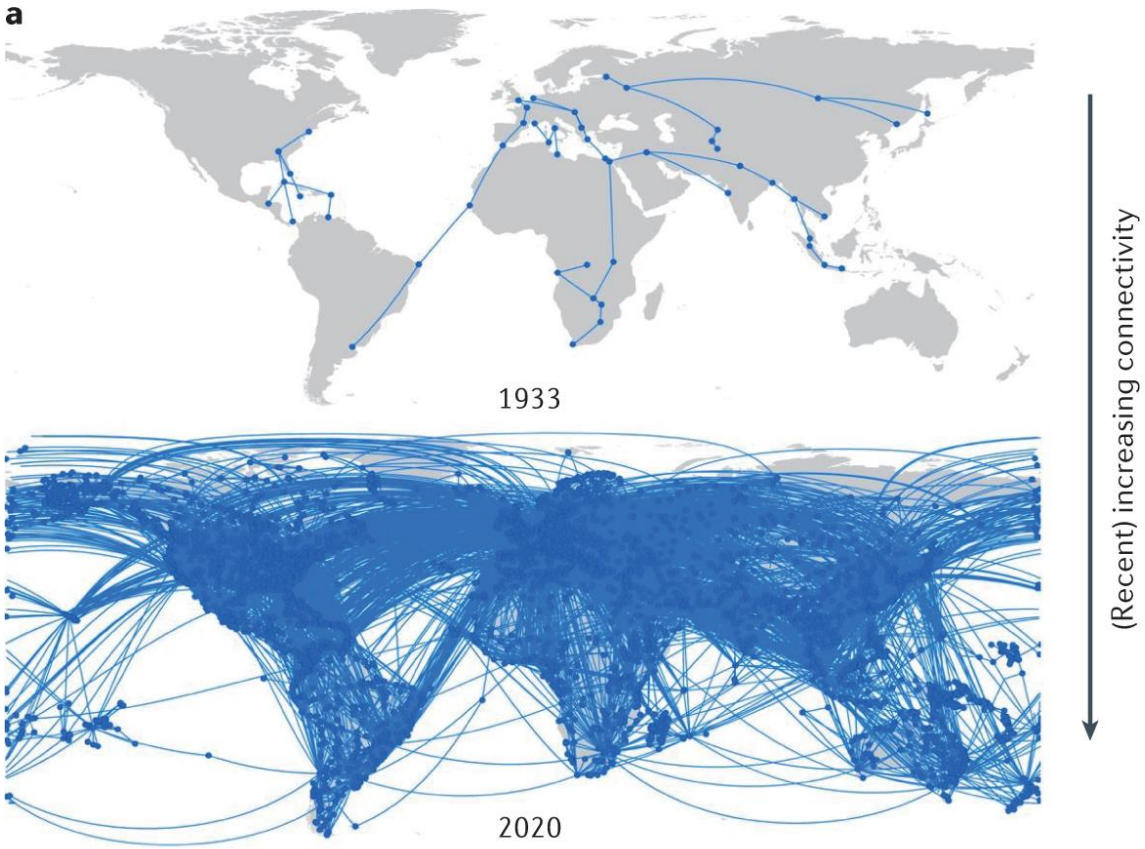
- Diseases that have not previously occurred in human hosts, have occurred in humans before but only infected small/isolated populations, or have occurred throughout time but have only recently been recognized as distinct diseases.
- Increasing global connectivity and climate change increase the chance for novel pathogen emergence and spread.
- Existing diagnostic methods may fail to detect an emerging pathogen.



Baker, R.E., et al. "Infectious disease in an era of global change". 2022. https://doi.org/10.1038/s41579-021-00639-z

Novel and Emerging Infectious Diseases

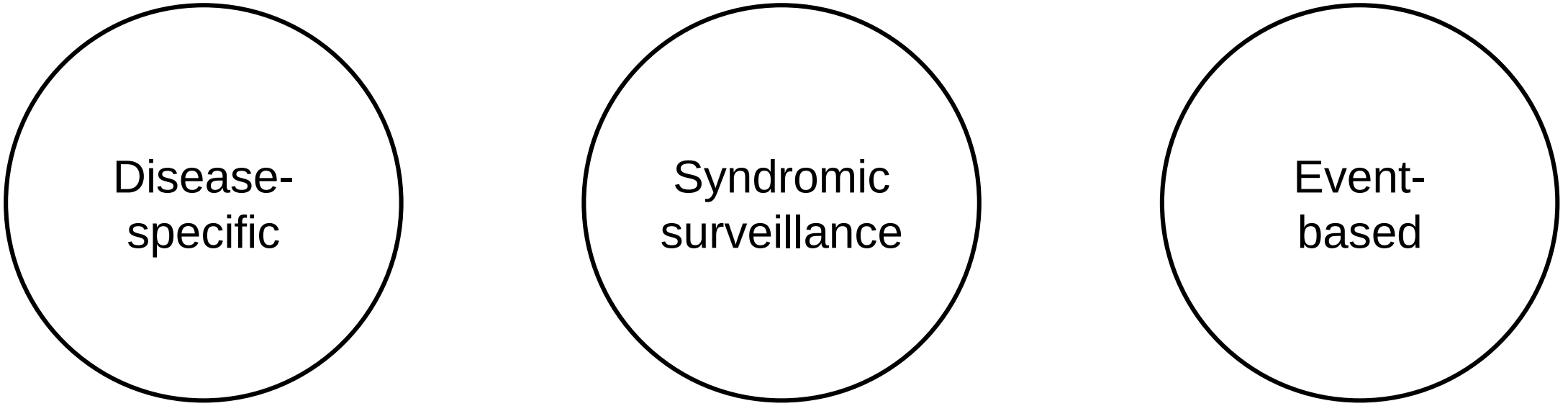
- Diseases that have not previously occurred in human hosts, have occurred in humans before but only infected small/isolated populations, or have occurred throughout time but have only recently been recognized as distinct diseases.
- Increasing global connectivity and climate change increase the chance for novel pathogen emergence and spread.
- Existing diagnostic methods may fail to detect an emerging pathogen.

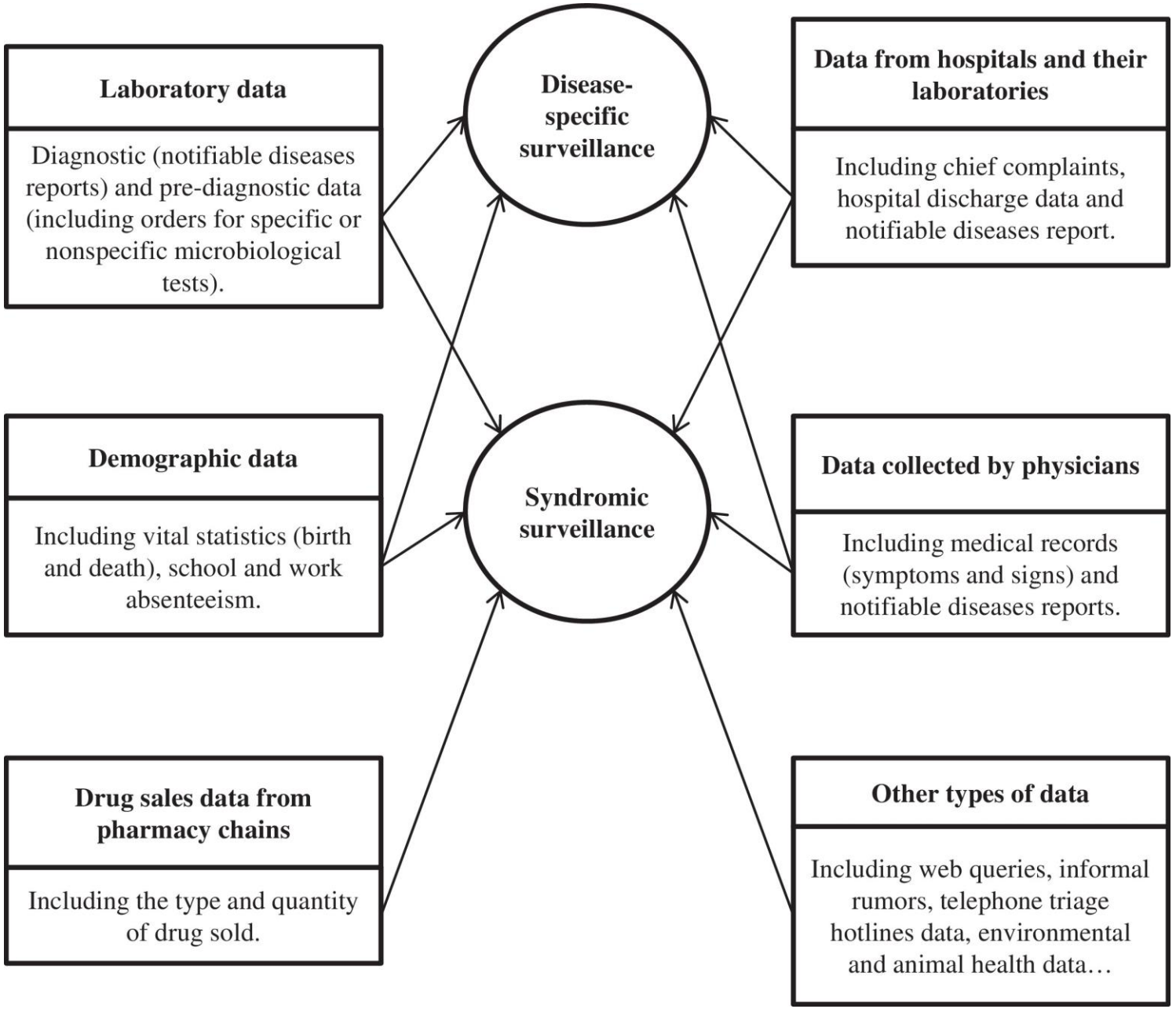


Baker, R.E., et al. "Infectious disease in an era of global change". 2022. https://doi.org/10.1038/s41579-021-00639-z

Infectious Disease Surveillance

- Describe the current burden and epidemiology of disease
- Monitor disease trends
- Identify outbreaks and new pathogens





Disease-specific: selection of diseases or syndromes

Syndromic surveillance: non-specific health indicators

Both involve structured data

Abat C, et al. "Traditional and syndromic surveillance of infectious diseases and pathogens". 2016. https://doi.org/10.1016/j.ijid.2016.04.021

Event-based surveillance

Table 2: Examples of Global Public Health Intelligence Network's successes in providing early warning signals for emerging communicable diseases

Disease	Date of first signal detected by GPHIN	Country where signal was detected	Type of source (Language of source)	Description of signal	Date of first report in the WHO Disease Outbreak News	Date of WHO declaration as a PHEIC	Date of first case confirmed in Canada
COVID-19 pandemic	December 31, 2019	China	International media reports (English)	Cases of viral pneumonia of unknown origin in Wuhan, China	January 5, 2020	January 30, 2020	January 25, 2020
2022 mpox outbreak	May 7, 2022	United Kingdom	Government health notice (English)	Confirmed case of mpox in London, England	May 16, 2022	July 23, 2022	May 19, 2022

Uses unstructured data (e.g., news reports) to provide early warning of new diseases

Norzin T, et al. "Event-based surveillance: Providing early warning for communicable disease threats". 2023. https://doi.org/10.14745/ccdr.v49i23a01

Event-based surveillance

Table 2: Examples of Global Public He
emerging communicable diseases

Disease	Date of first signal detected by GPHIN	Coun whe sign wa detec					
COVID-19 pandemic	December 31, 2019	China	reports (English)	unknown origin in Wuhan, China			
2022 mpox outbreak	May 7, 2022	United Kingdom	Government health notice (English)	Confirmed case of mpox in London, England	May 16, 2022	July 23, 2022	May 19, 2022

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

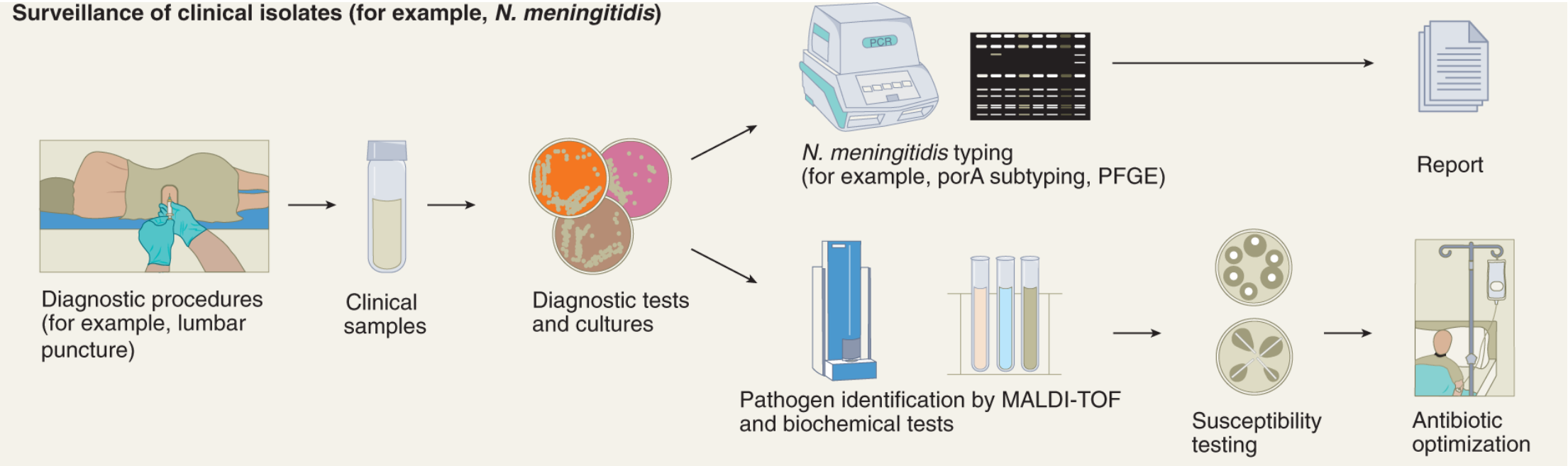
Further investigation is required to confirm, identify, and characterize the cause of the disease.
May include laboratory-based tests and sequencing.

Norzin T, et al. “Event-based surveillance: Providing early warning for communicable disease threats”. 2023. https://doi.org/10.14745/ccdr.v49i23a01

Infectious Disease Surveillance

- Describe the current burden and epidemiology of disease
- Monitor disease trends
- Identify outbreaks and new pathogens

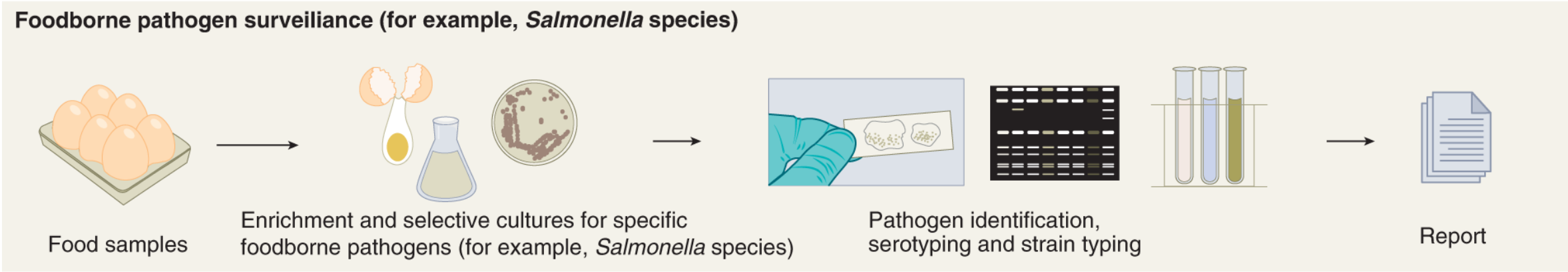
Laboratory-Based Pathogen-specific Surveillance



Clinical

Ko, K.K.K., et al. "Metagenomics-enabled microbial surveillance". 2022. https://doi.org/10.1038/s41564-022-01089-w

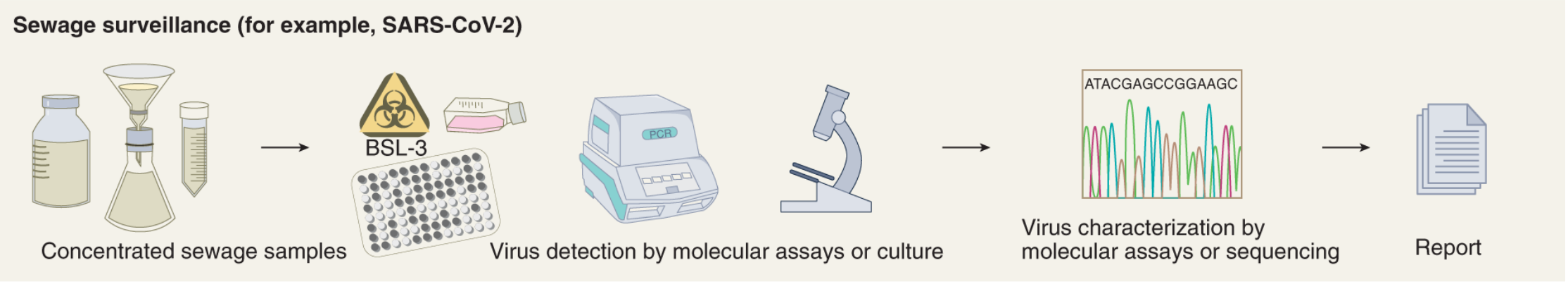
Laboratory-Based Pathogen-specific Surveillance



Foodborne

Ko, K.K.K., et al. "Metagenomics-enabled microbial surveillance". 2022. https://doi.org/10.1038/s41564-022-01089-w

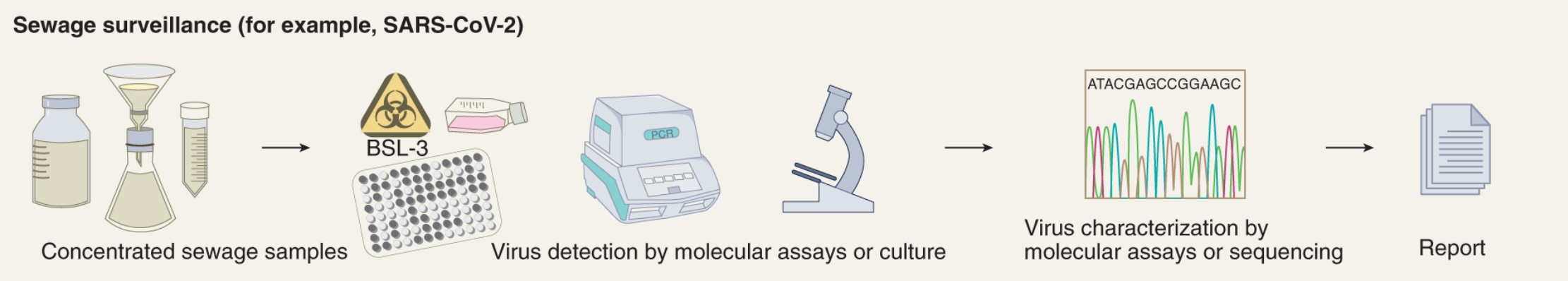
Laboratory-Based Pathogen-specific Surveillance



Environmental

Ko, K.K.K., et al. "Metagenomics-enabled microbial surveillance". 2022. https://doi.org/10.1038/s41564-022-01089-w

Laboratory-Based Pathogen-specific Surveillance



Environmental

These all require laboratory-based testing techniques specific to a particular pathogen.

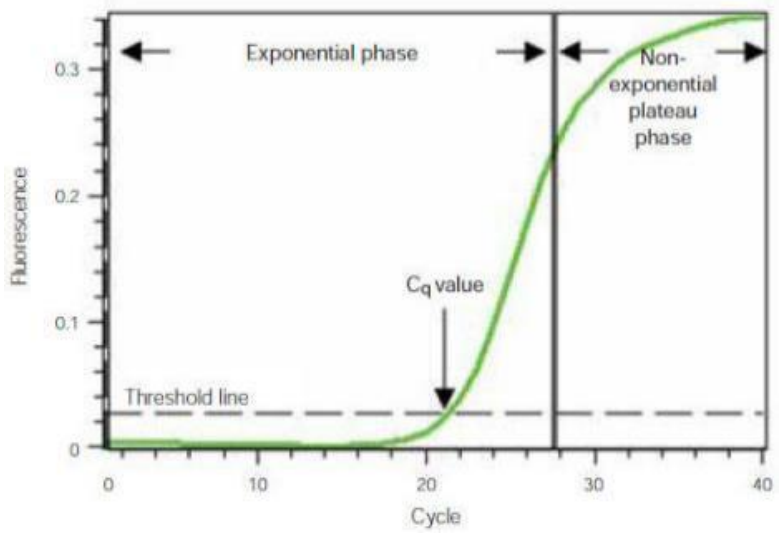
Ko, K.K.K., et al. "Metagenomics-enabled microbial surveillance". 2022. https://doi.org/10.1038/s41564-022-01089-w

Traditional Laboratory-Based Pathogen Detection

- Nucleic acid amplification
- Serotyping
- Culture
- Proteomics
- Microscopy

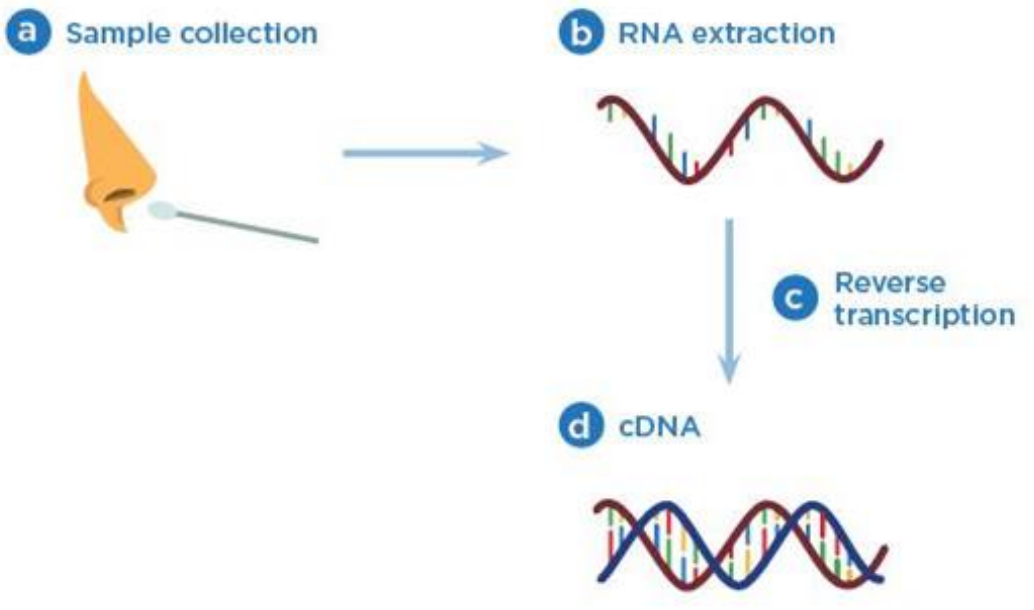
Nucleic Acid Amplification-based techniques

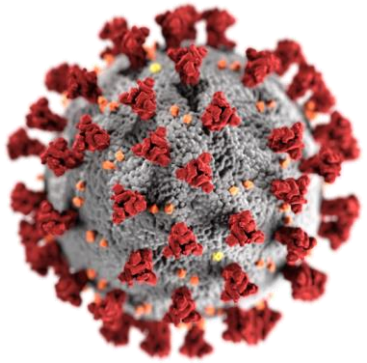
- Developed for subtyping pathogens based on genetic amplification and analysis of “signatures” contained in their genome

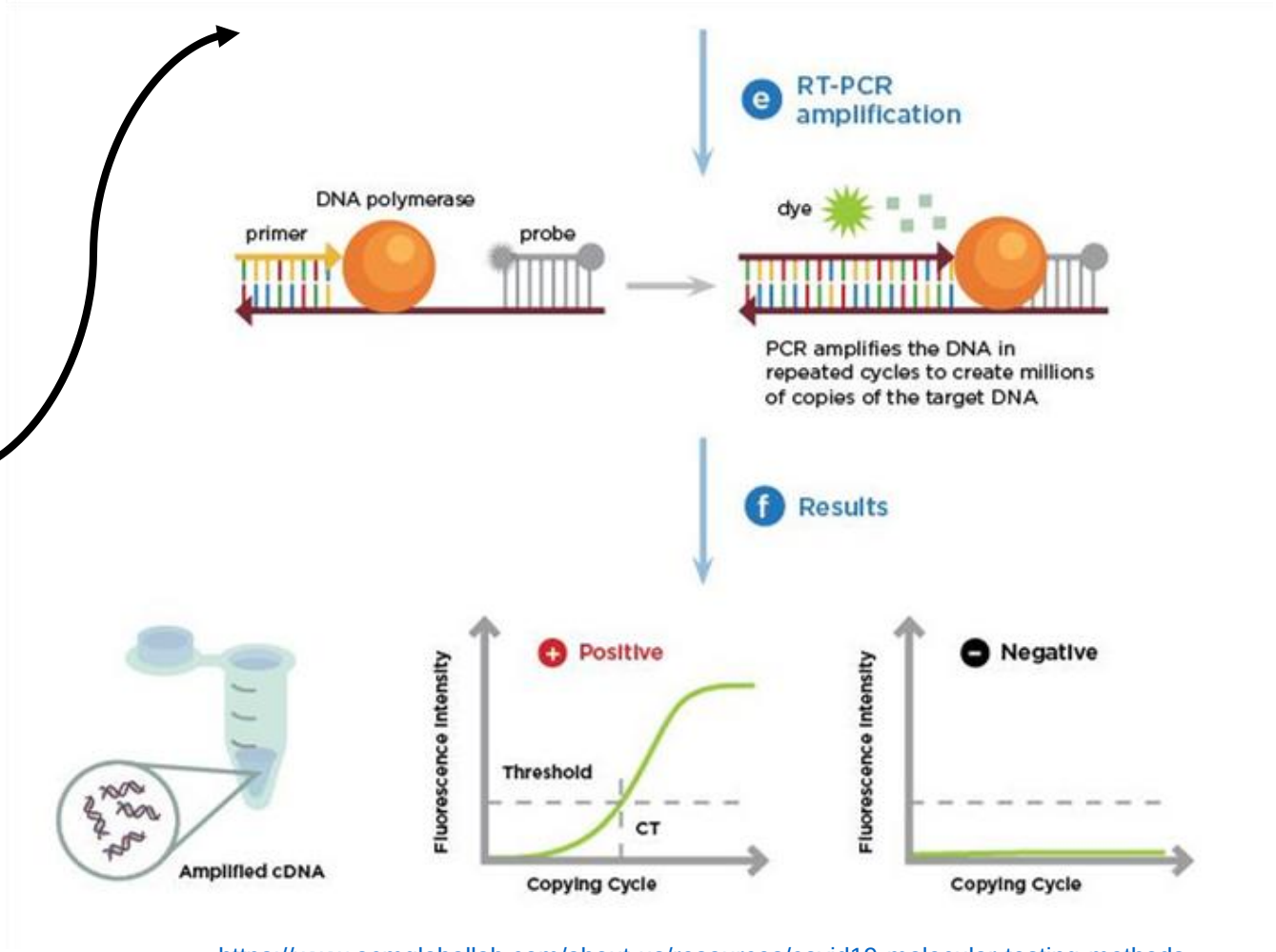


https://www.cbc.ca/news/canada/british-columbia/covid19-minorities-health-bc-canada-1.5801777

Real-time Quantitative PCR (RT-qPCR)



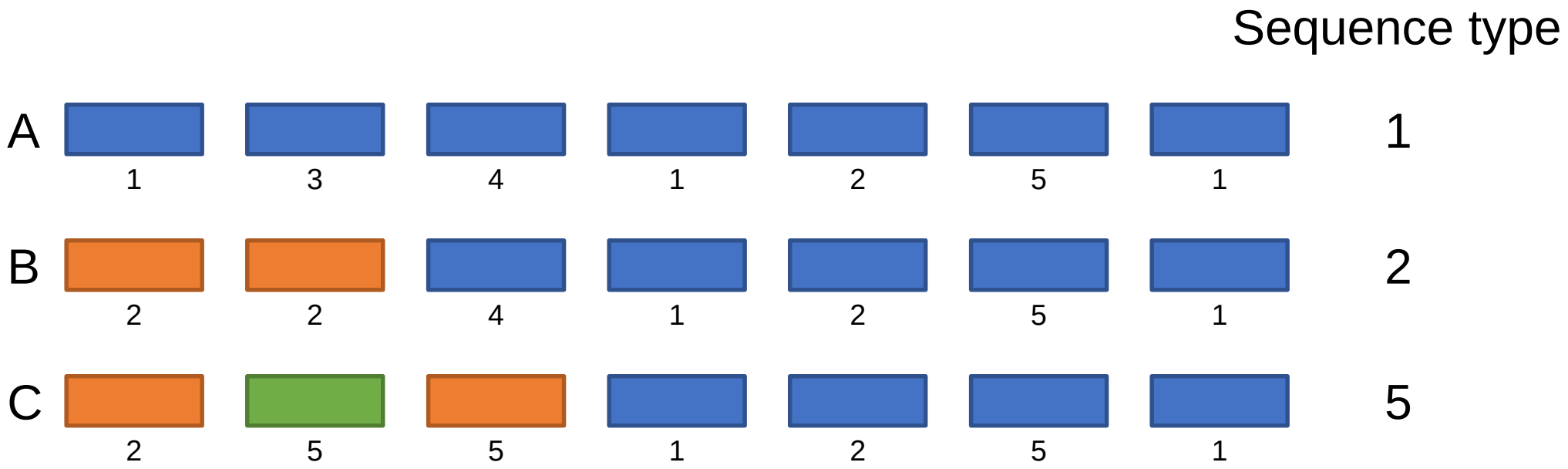




https://www.acmgloballab.com/about-us/resources/covid19-molecular-testing-methods

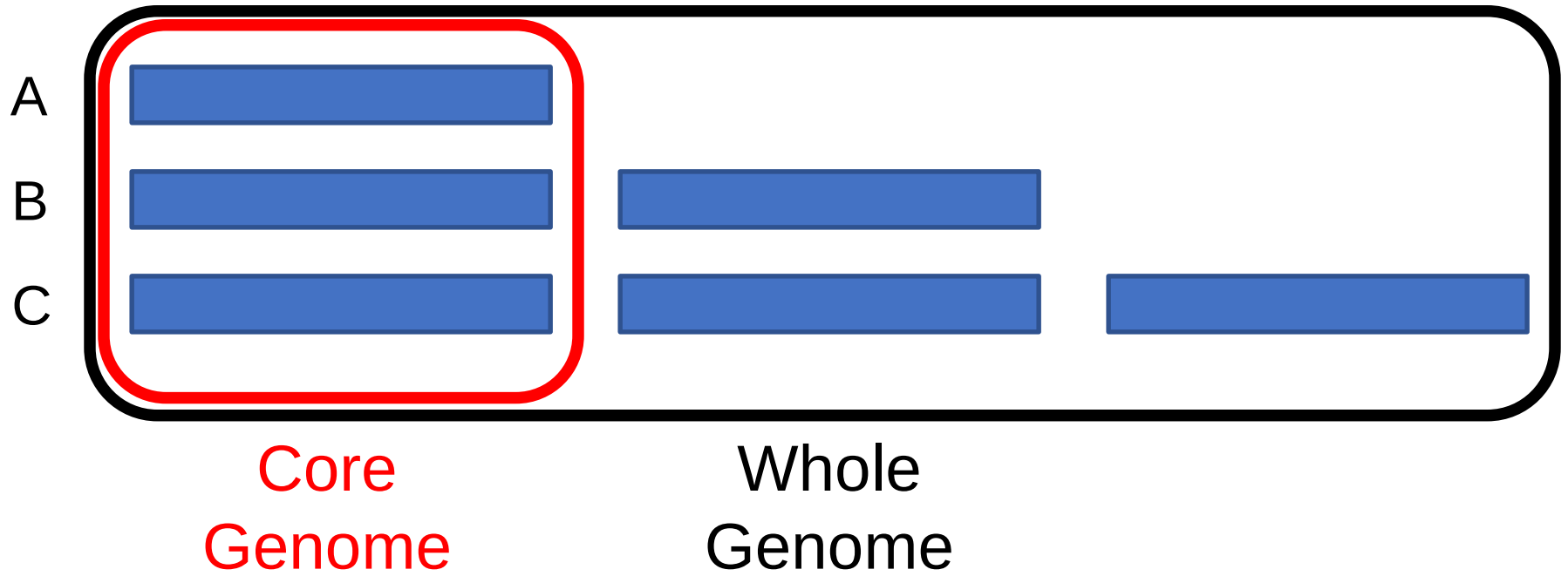
Multi-locus sequence typing

- Based on 6 or 7 genes, different alleles assigned a number



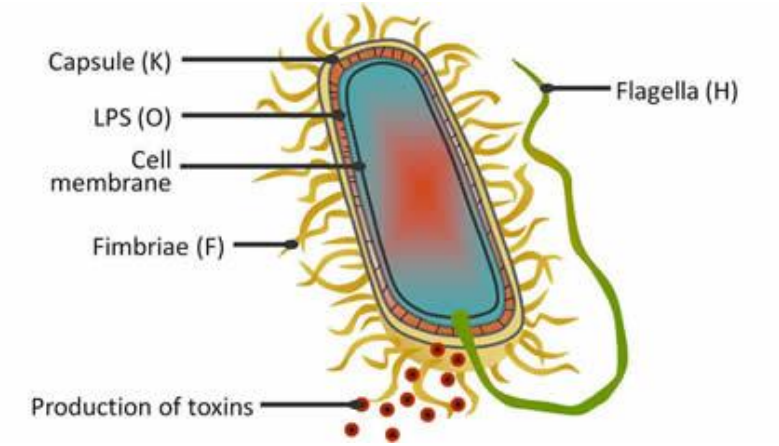
cg/wgMLST

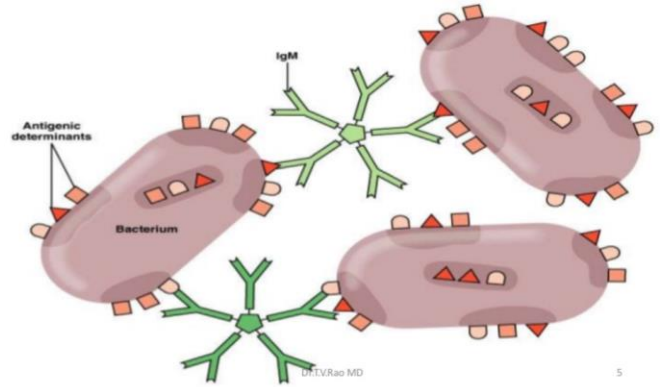
- Extension of MLST to 100s or 1000s of genes
  - **cgMLST**: core genome MLST
  - **wgMLST**: whole genome MLST

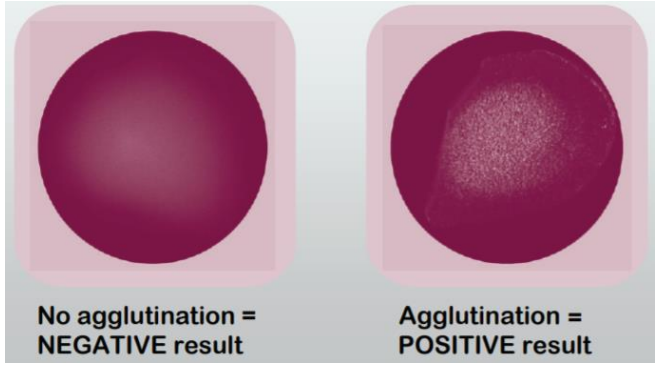


Serotyping

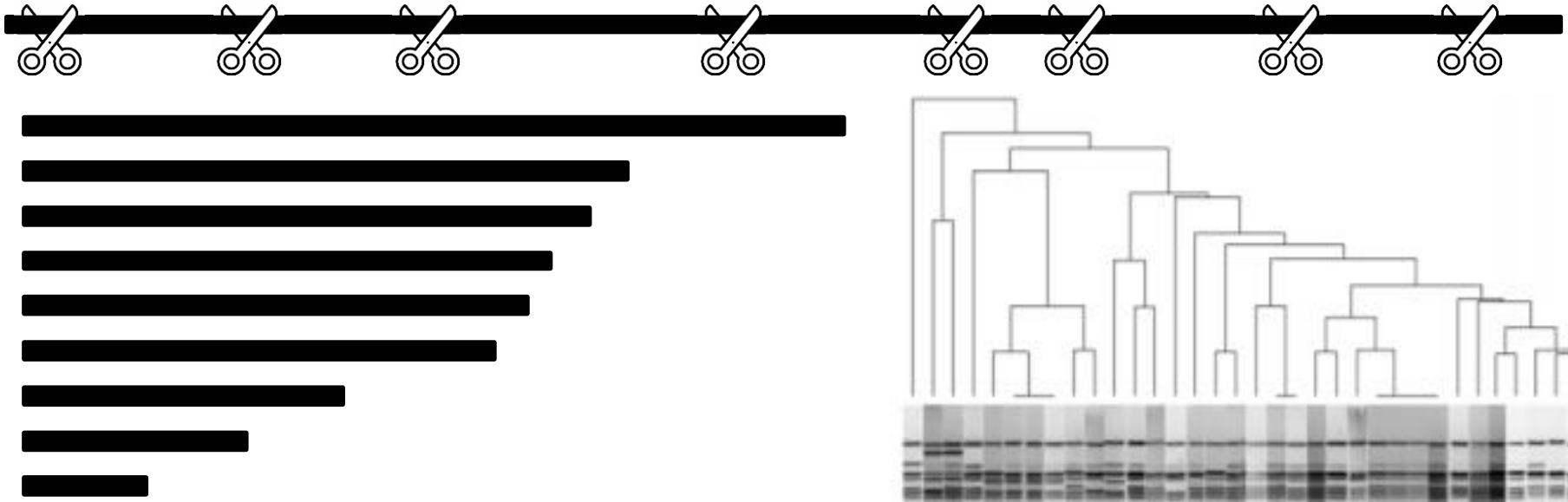
- **Serotyping** is a subtyping method commonly used for bacteria and viruses. The method uses cell surface molecules (also known as antigens) produced by the organism.







Restriction Digestion-Based Methods

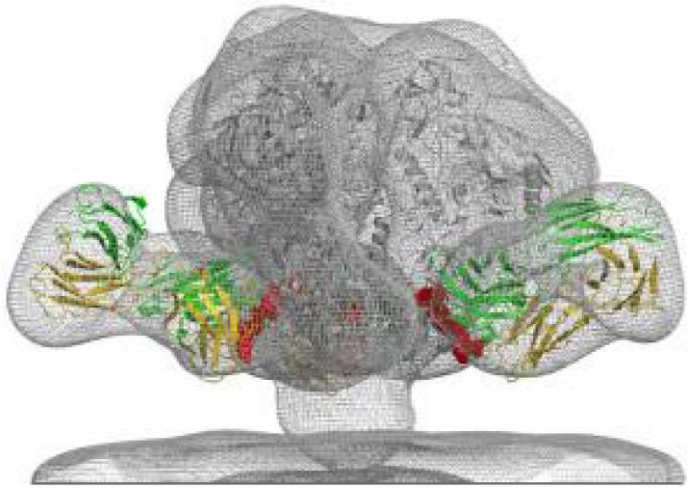


Restriction enzymes cut DNA at specific locations to create fragments of various sizes, sorted by size on a gel.

- **Pulsed-Field Gel Electrophoresis (PFGE)**
- **Ribotyping**

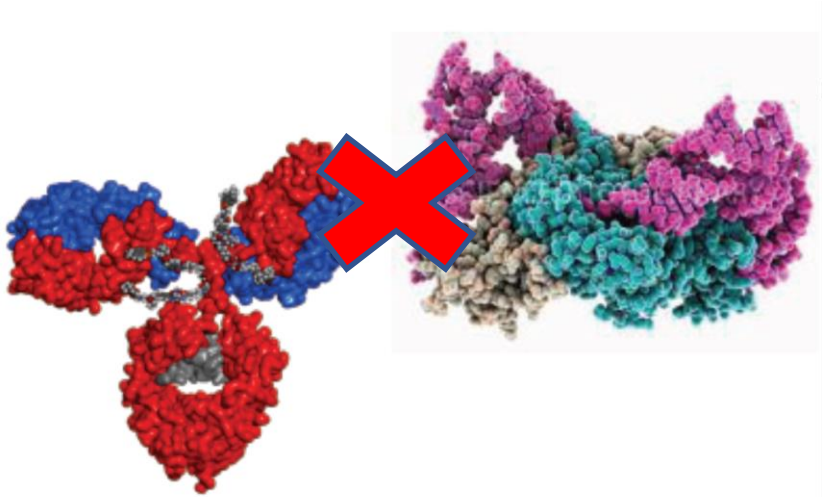
Advantages and Disadvantages of Traditional Diagnostic Methods For Pathogen Detection

- Advantages:
  - Specificity
  - If it's there, you'll detect it.



```
TAFQEALDAAGDKLVVVDFSATmCGRCKMIKPFFHSLSEKYSNVIFLEVDV
:::::::::::::::::::::-:::::::::::::::::::::::::::::
TAFQEALDAAGDKLVVVDFSATWCGPCKMIKPFFHSLSEKYSNVIFLEVDV
```

- Disadvantages:
  - Sensitivity
  - If it's there, but it has evolved, you might miss it.



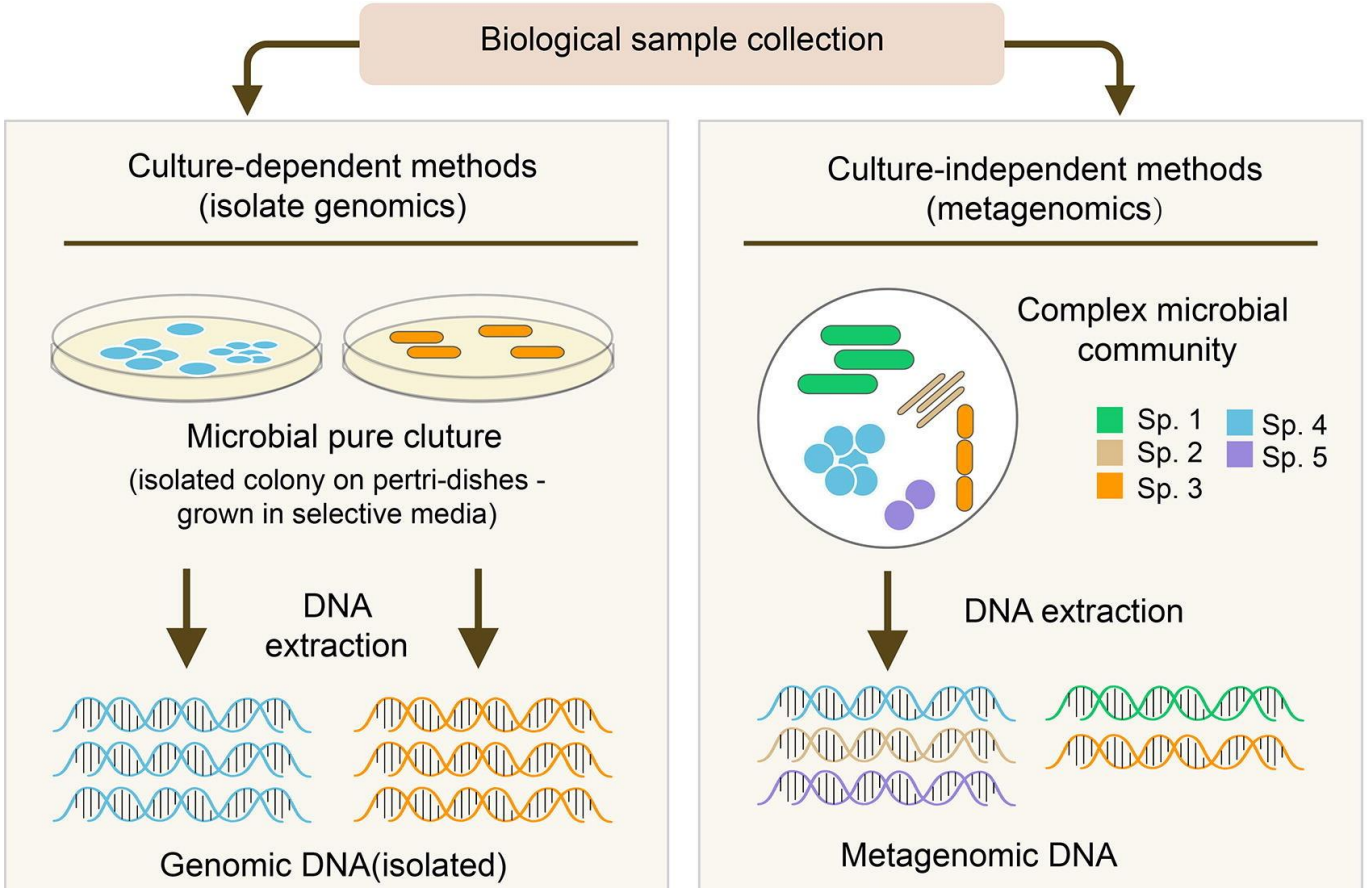
```
TAFQEALDAAG--DKLVVlDFSATWgGRCKMIrPFFHSLSEKISNVILPFLEVDV
:::::::::::--::::-:::::::-:::::::-:::::::::-::::--::::::
TAFQEALDAAGMMDKLVVVDFSATWCGPCKMIKPFFHSLSEKYSNVI--FLEVDV
```

Shotgun Metagenomics for Novel and Emerging Pathogen Detection

- Shotgun metagenomics is a genome-sequencing technique that provides an “unbiased” survey of nucleic acid content.
- Nucleic acid can be DNA or RNA.
- Clinical metagenomic specimens contain host nucleic acid plus microbial nucleic acid (commensal plus pathogen).
- Metagenomic samples can also contain contaminating DNA from sources external to the sample.

Comparison of culture-dependent genomics to metagenomics

Isolate genomics

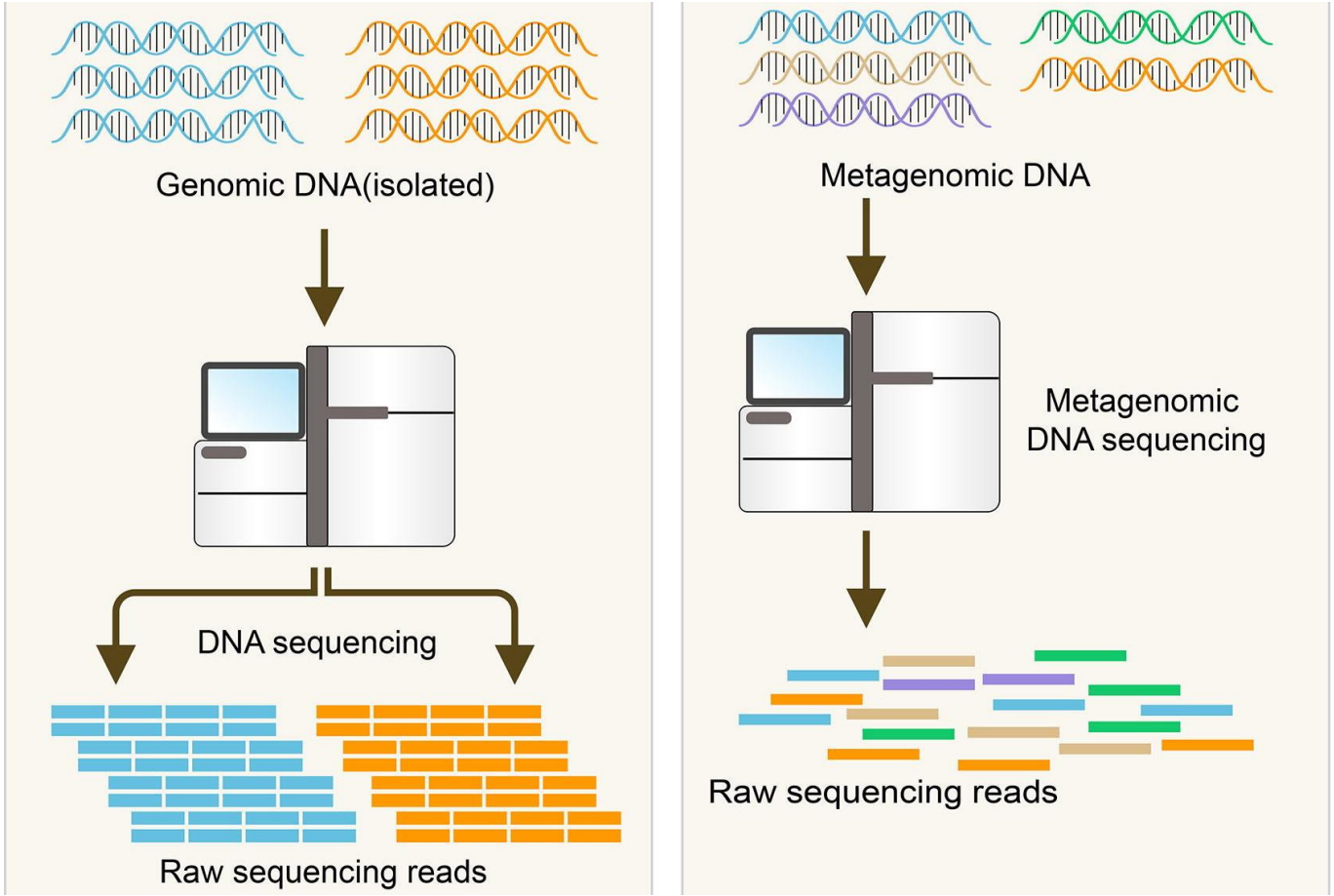


Metagenomics

Yang C, et al. 2021. "A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data". https://doi.org/10.1016/j.csbj.2021.11.028

Comparison of culture-dependent genomics to metagenomics

Isolate genomics

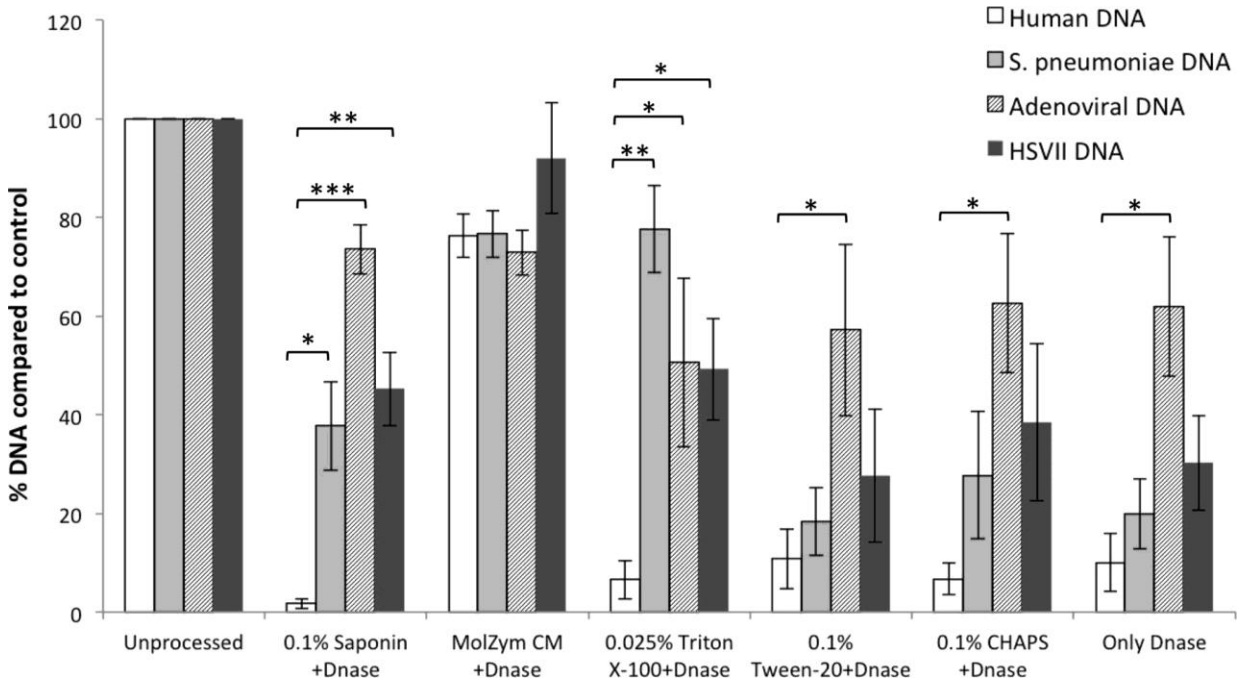


Metagenomics

Yang C, et al. 2021. "A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data". https://doi.org/10.1016/j.csbj.2021.11.028

Host Reduction

- The relative abundance of host nucleic acid varies widely with specimen type (e.g., < 5 % for feces; > 99 % for cerebrospinal fluid).
- Wet-lab and computational methods can be used to remove host nucleic acids.
- Wet Lab:
  - Host nucleic acid is reduced, or microbial nucleic acid is enriched.
  - Works best with samples with high microbial content.
  - Many different methods are available:
    - CpG island hybridization
    - rRNA depletion
    - Poly-A selection
    - Selective host cell lysis and DNA degradation
- Computational Approaches
  - Map to host genome and remove.

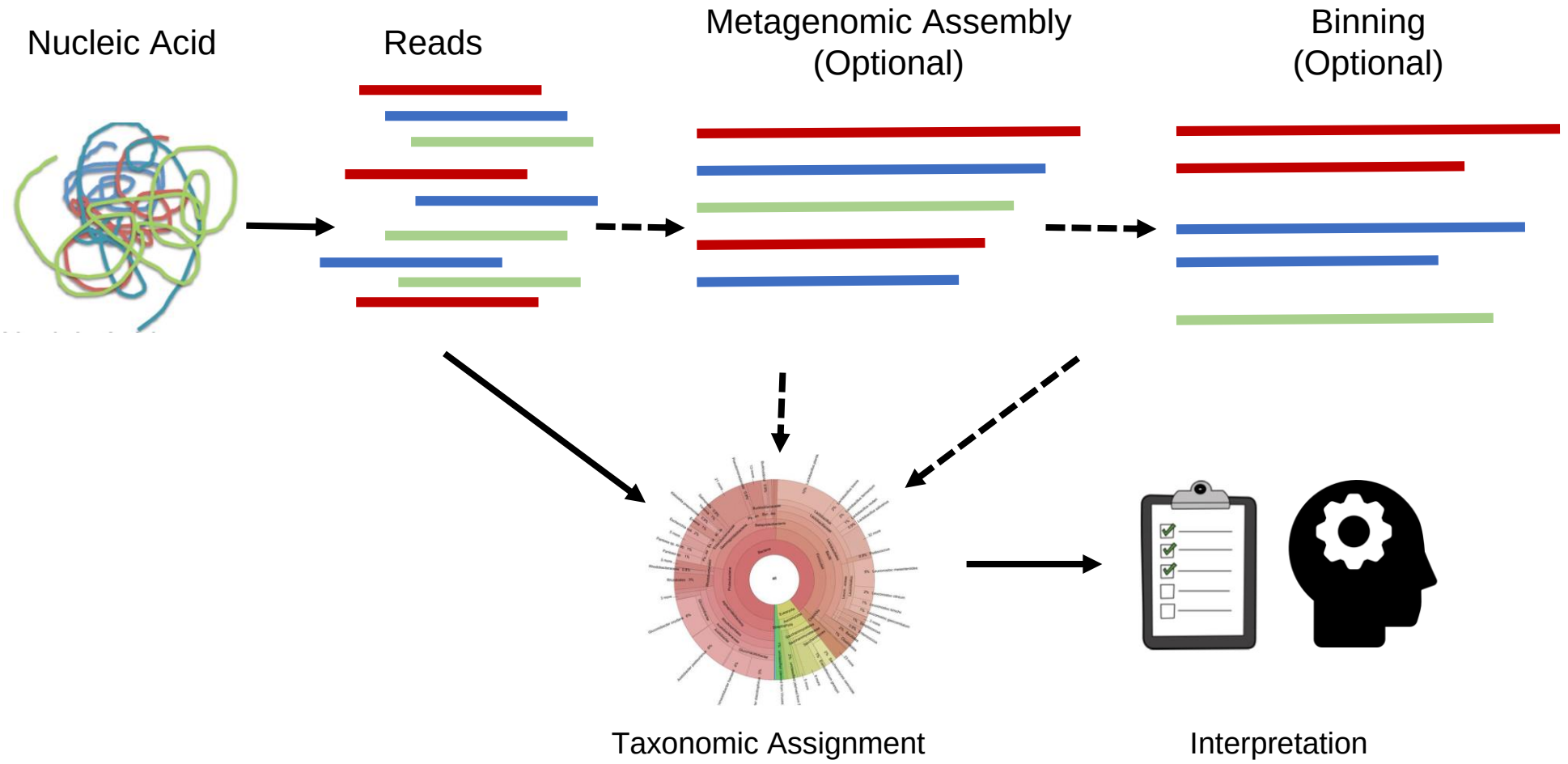


Hasan MR, et al. "Depletion of Human DNA in Spiked Clinical Specimens for Improvement of Sensitivity of Pathogen Detection by Next-Generation Sequencing". 2016. https://doi.org/10.1128/JCM.03050-15

Contamination

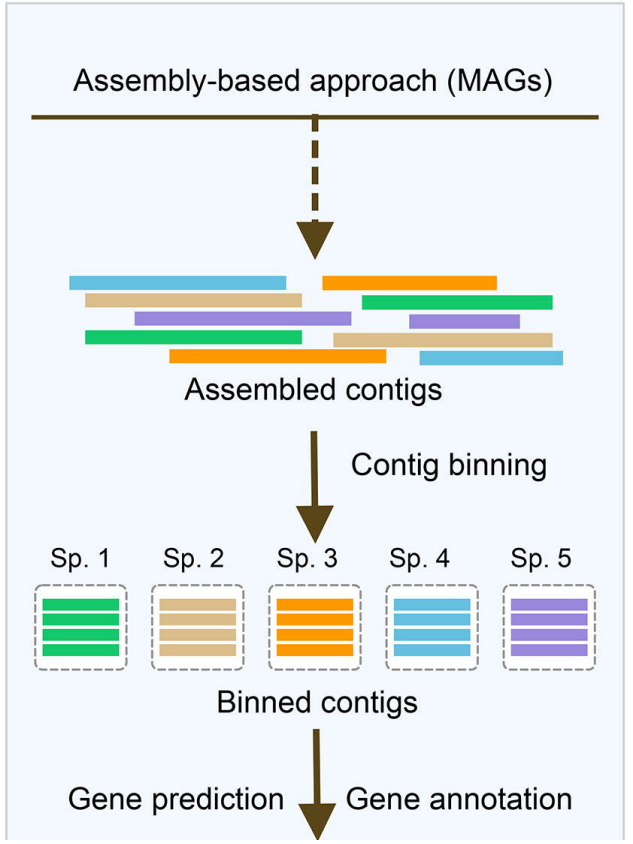
- Nucleic acid external to the sample can be introduced into the sample at all stages of sample preparation:
  - Collection
  - Extraction
  - Library preparation
- Contamination can come from multiple sources:
  - Lab reagents
  - Lab equipment
  - Lab coats
  - Lab workers
- Negative (“no template”) controls (blanks) are used to identify contamination.

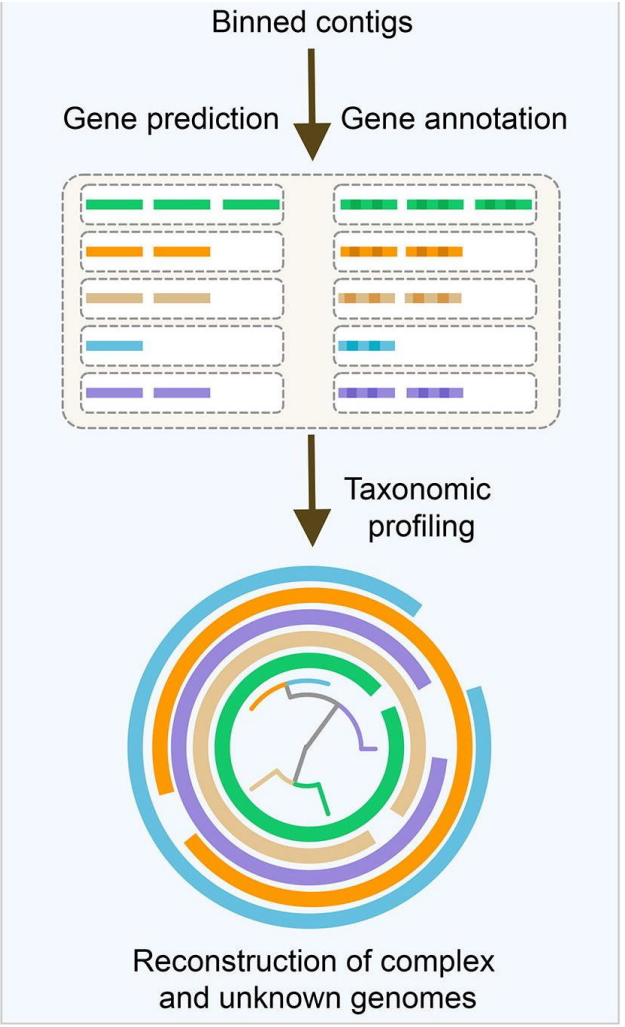
Overall process for Pathogen Detection from Metagenomic Data



Metagenomic assembly and binning (optional steps)

- Metagenomic Assembly
  - Megahit
  - MetaSPAdes
  - Many others
- Binning
  - Groups contigs (or reads) based on shared characteristics
    - Tetranucleotide frequencies
    - Abundances
    - Codon usage
  - Forms a Metagenome Assembled Genome (MAG)
- Use for taxonomic assignment or phylogenetic analysis



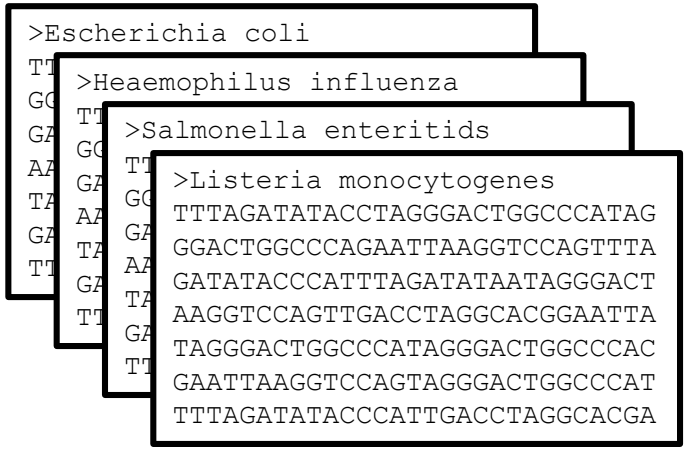


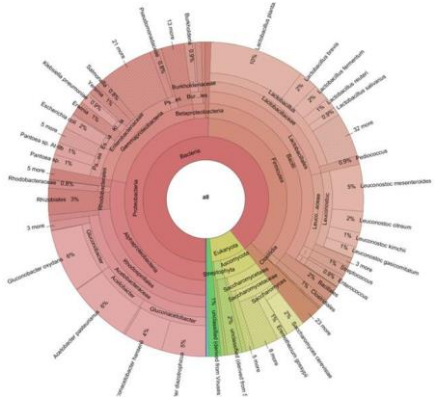
Yang C, et al. 2021. "A review of computational tools for generating metagenome-assembled genomes from metagenomic sequencing data". https://doi.org/10.1016/j.csbj.2021.11.028

Taxonomic Assignment

- Taxonomic assignment is the process of assigning reads (or contigs) to a "taxon."
- Taxonomic assignment can occur at different taxonomic ranks:
  - Phylum > Class > Order > Family > Genus > Species > Subspecies
- Assignment is typically performed by comparing the read to a reference database of known taxa.

TTGACCTGGAGG



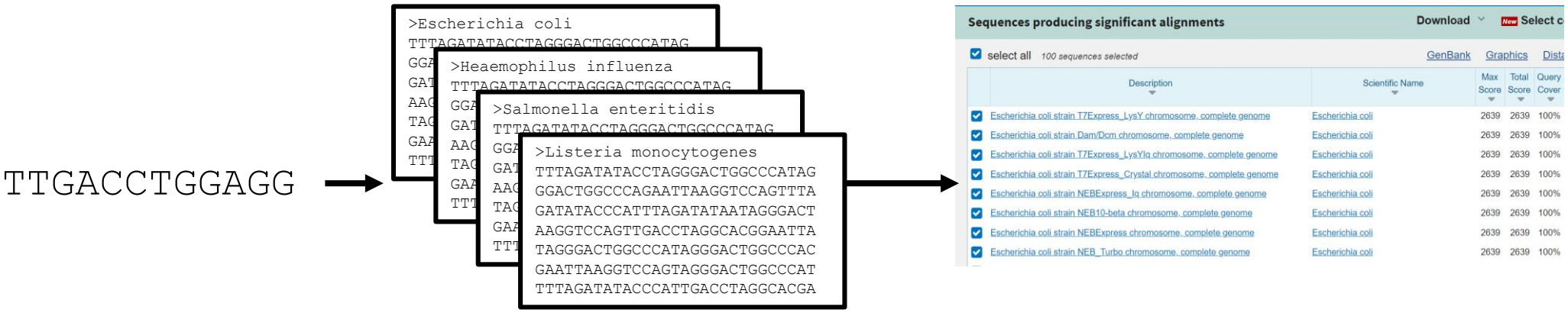


Which taxonomic assignment method is best?

- Assignment methods should be
  - Fast
  - Sensitive
  - Specific

Assignment using BLAST

- BLAST is a fast sequence alignment program.
- BLAST parameters can be adjusted to fine tune specificity and sensitivity.
- Metagenomic data can generate tens of millions of reads.
- BLAST is fast, but not fast enough for the job under most circumstances (for reads).
- There is the additional challenge of managing the data.



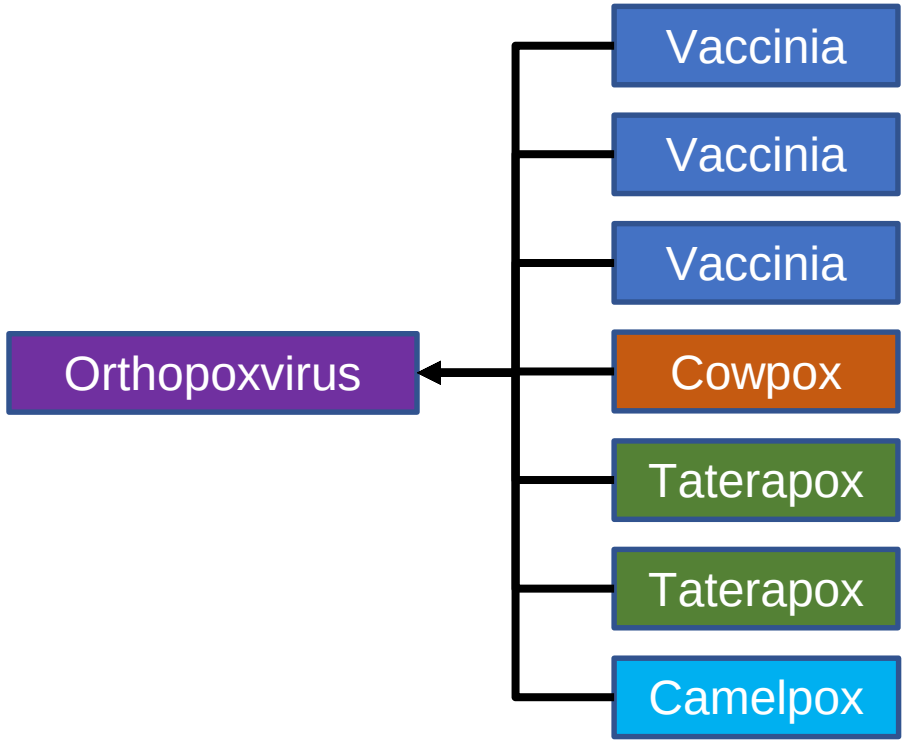
Assignment using BLAST

- How do you choose the correct BLAST hit?

Name	BLAST Score
Campylobacter	150
Arcobacter	150
Ehrlichia	140
Closridium botulinum	80
Methanococcus	20

Assignment using BLAST

- How do you choose the correct taxonomic rank?



Taxonomic Assignment With Kraken

Kraken is an ultrafast and highly accurate program for assigning taxonomic labels to metagenomic DNA sequences.

Method | Open Access | Published: 03 March 2014

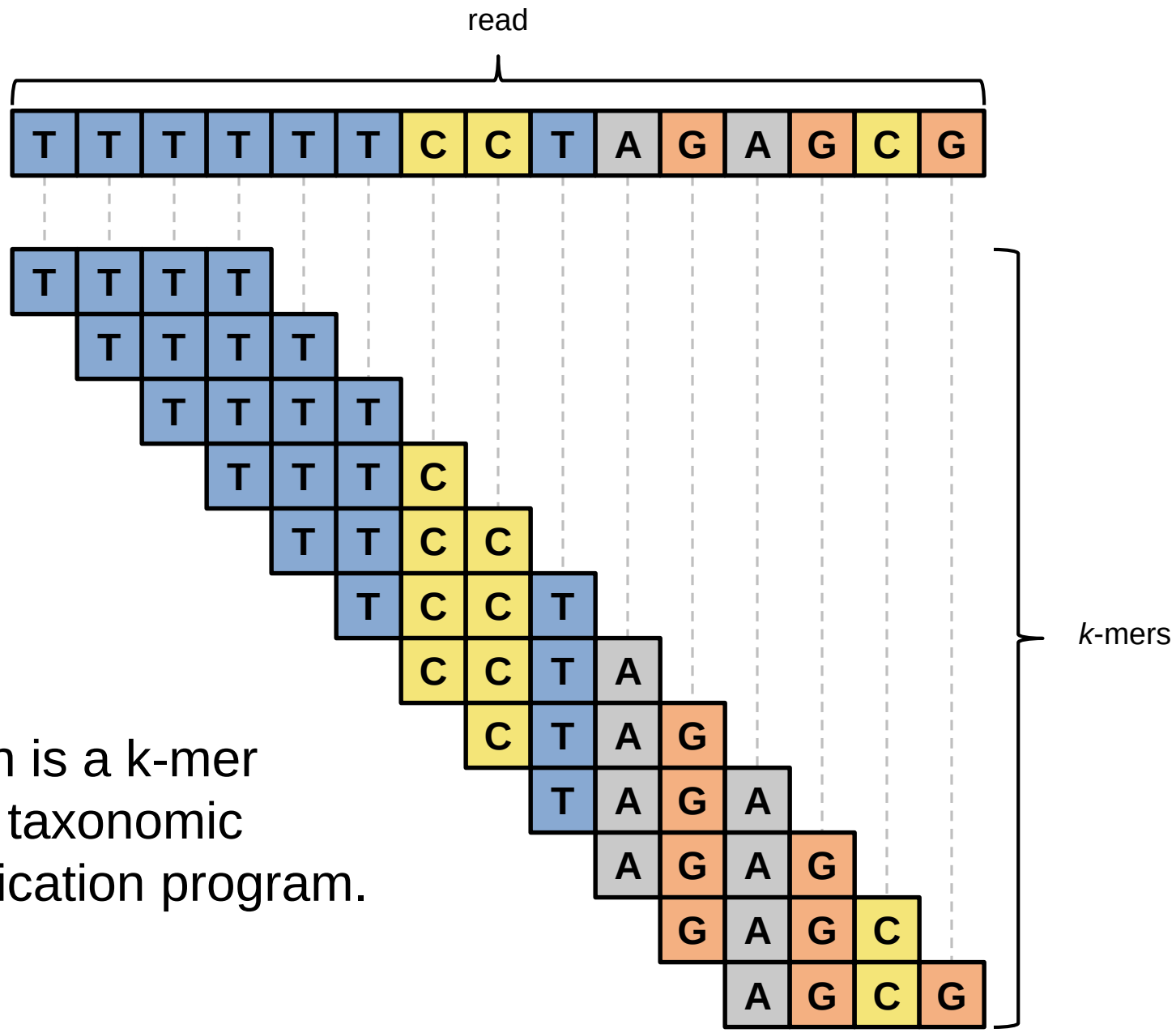
Kraken: ultrafast metagenomic sequence classification using exact alignments

Derrick E Wood & Steven L Salzberg

Short Report | Open Access | Published: 28 November 2019

Improved metagenomic analysis with Kraken 2

Derrick E. Wood, Jennifer Lu & Ben Langmead

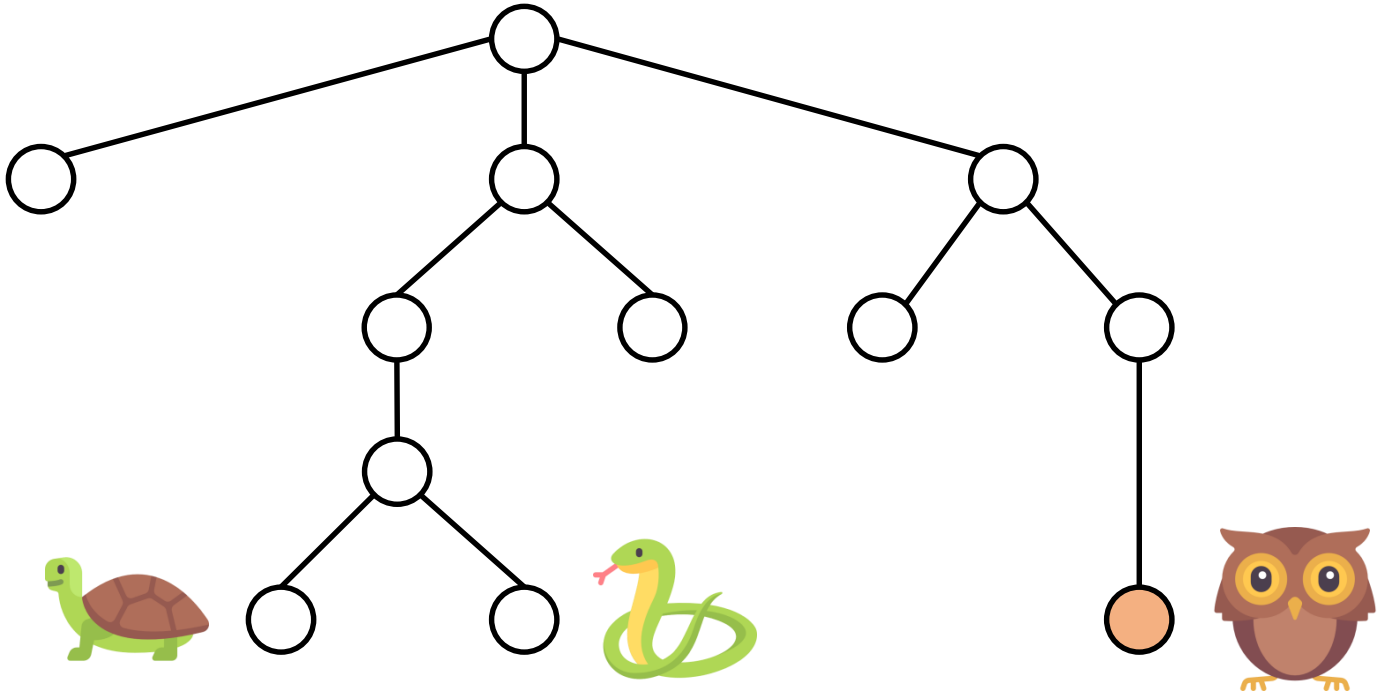


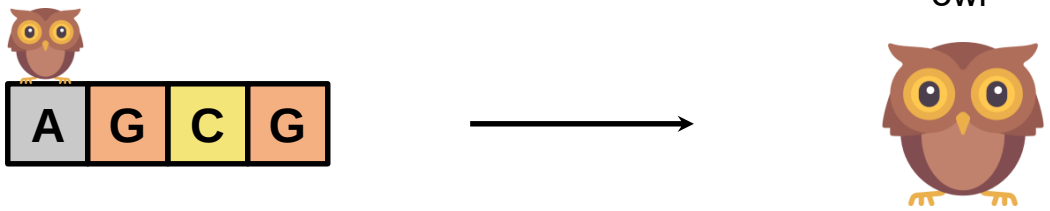
Kraken is a k-mer based taxonomic classification program.

Building a Kraken Database

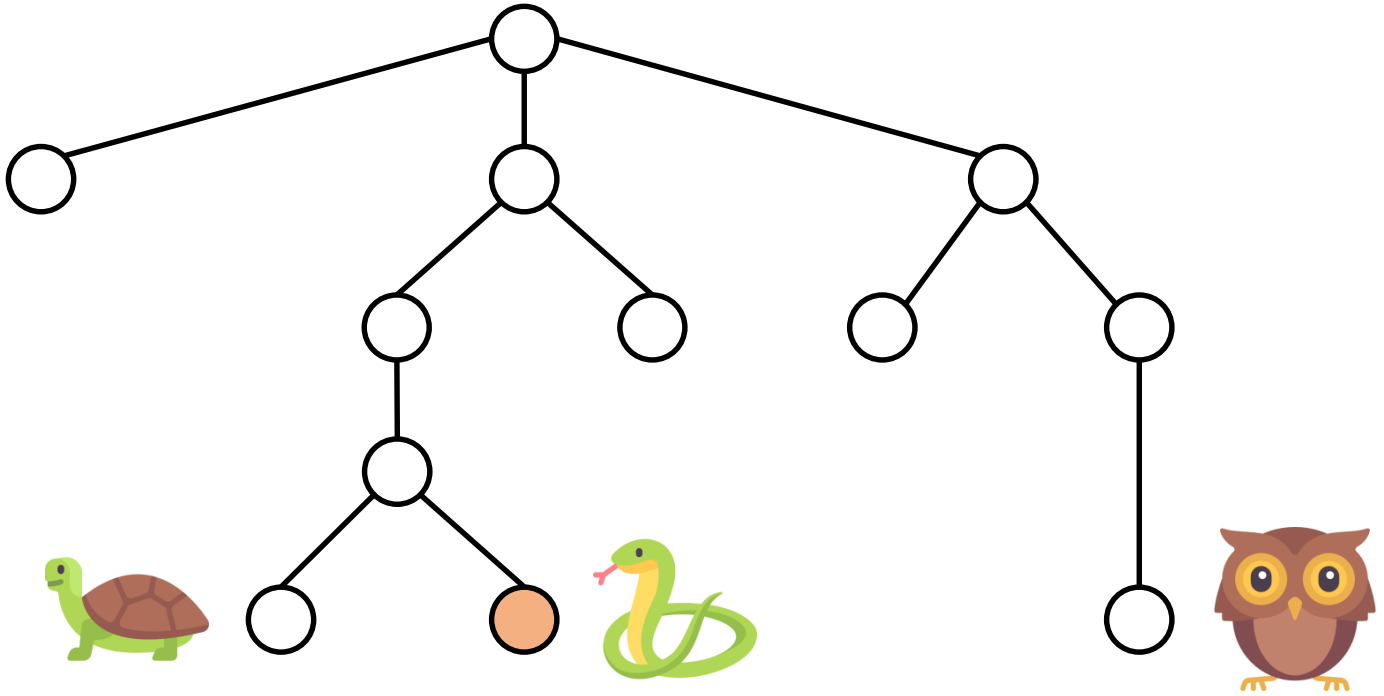
T C C T	T C C T	T T T T
A G C G	G G C C	G G C C
C T G A	A G A G	A G A G
T A A T	T A A T	T A A T

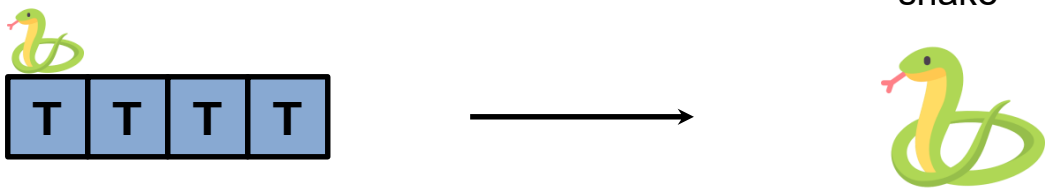
Building a Kraken Database



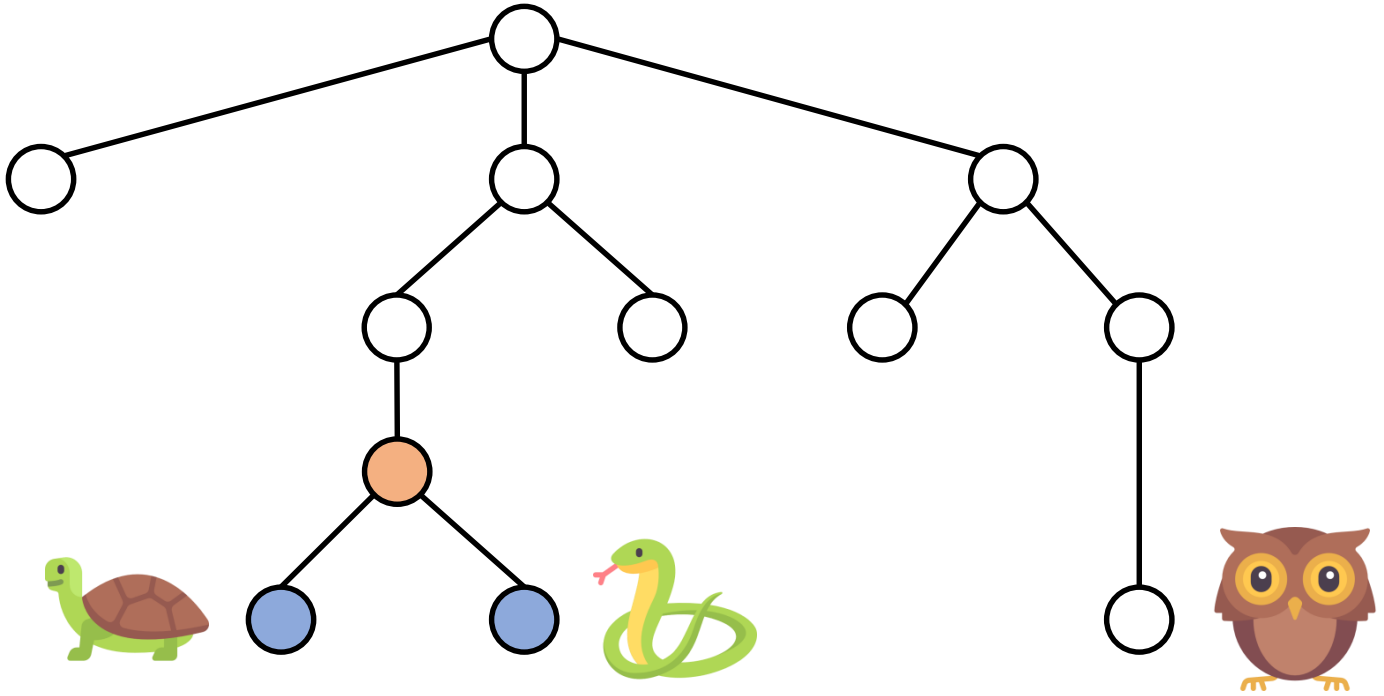


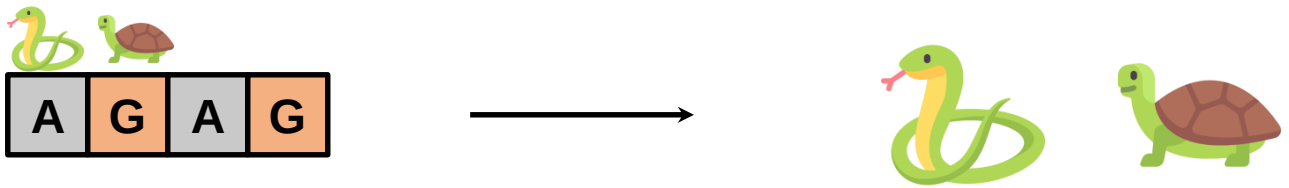
Building a Kraken Database



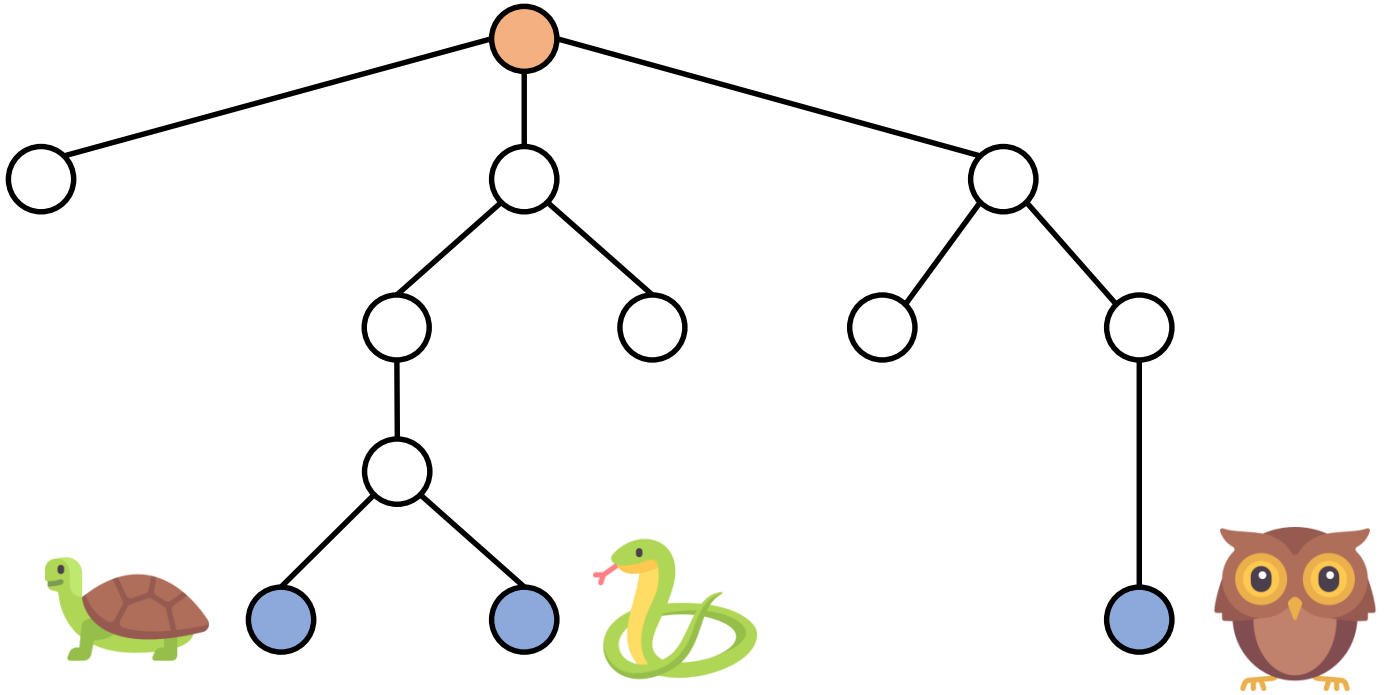


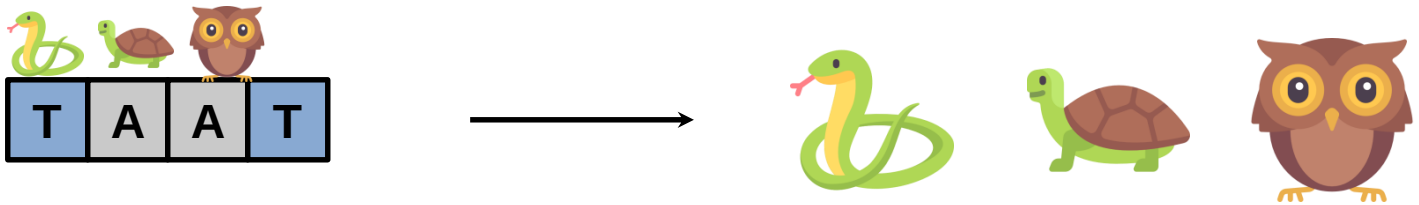
Building a Kraken Database





Building a Kraken Database

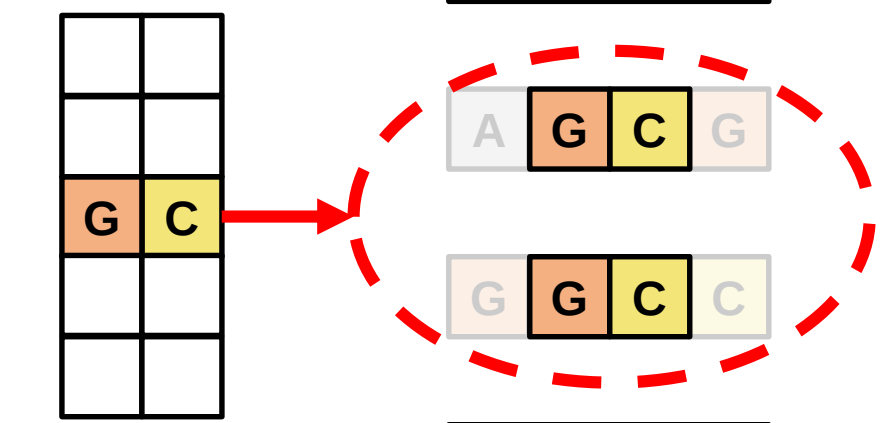




Database

k-mer	observations	classification
T C C T	🦉 🐢	“vertebrate”
T T T T	🐍	“snake”
A G C G	🦉	“owl”
G G C C	🐢 🐍	“reptile”
A G A G	🐢 🐍	“reptile”
C T G A	🦉	“owl”
T A A T	🦉 🐢 🐍	“vertebrate”

Database



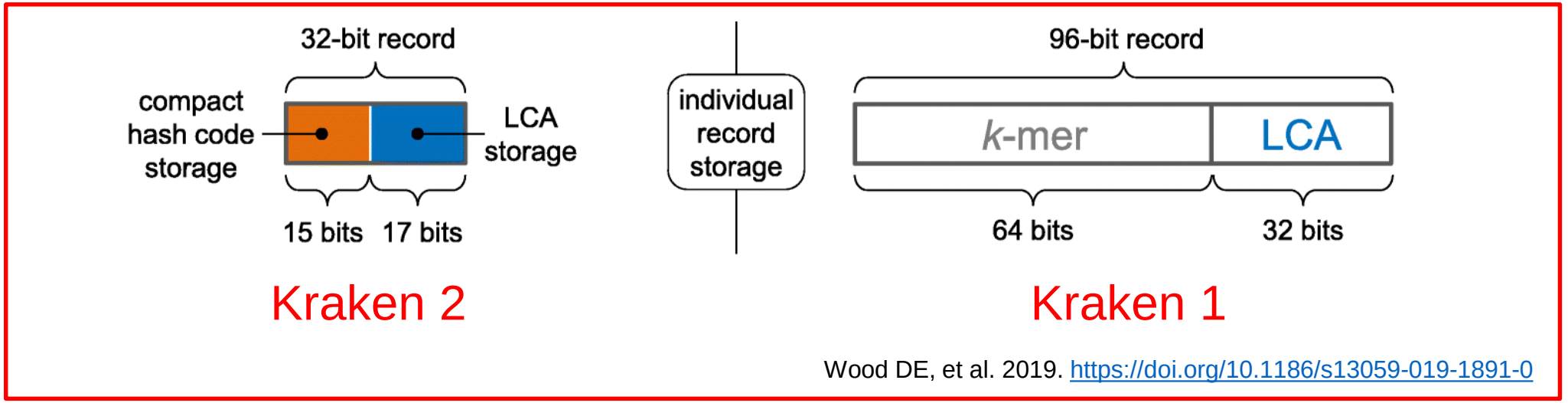
Minimizer

k-mer	observations	classification
T C C T	🦉 🐢	“vertebrate”
T T T T	🐍	“snake”
A G C G	🦉	“owl”
G G C C	🐢 🐍	“reptile”
A G A G	🐢 🐍	“reptile”
C T G A	🦉	“owl”
T A A T	🦉 🐢 🐍	“vertebrate”

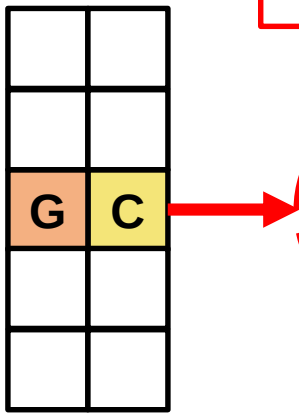
Kraken also groups k-mers by their common **minimizer** (an l-mer of the k-mer).

Used for more efficient searching through the database.

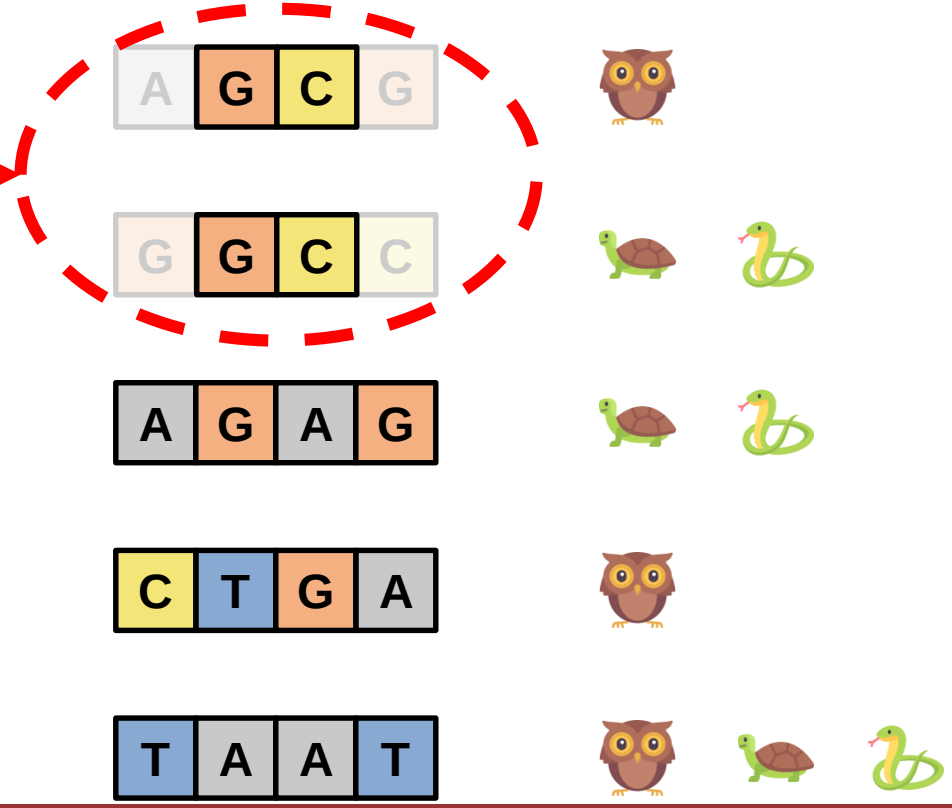
Database



Wood DE, et al. 2019. https://doi.org/10.1186/s13059-019-1891-0



Minimizer



"owl"

"reptile"

"reptile"

"owl"

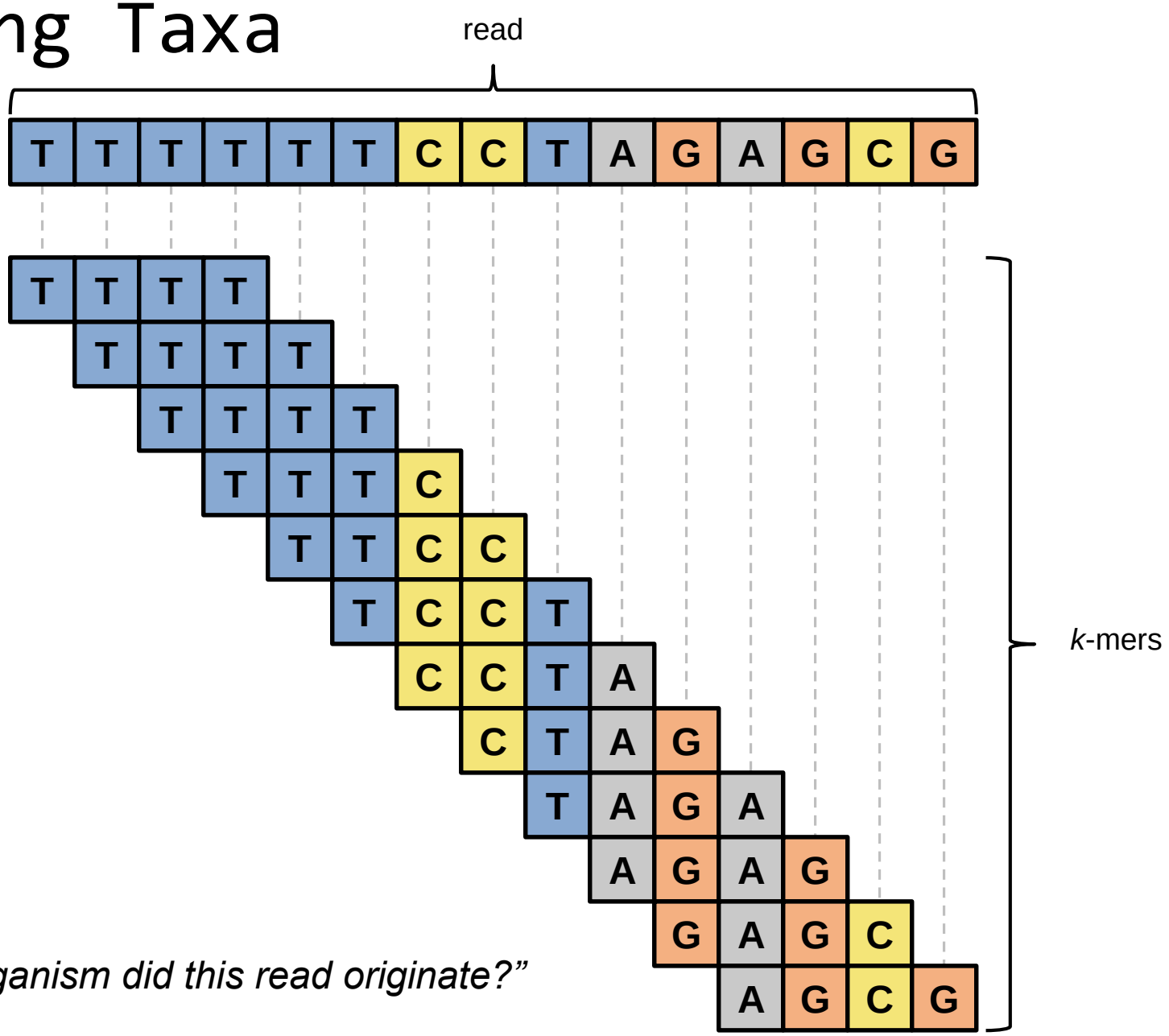
"vertebrate"

Kraken 2 reduces the database size by storing compact hashes derived from minimizers in the database.

Database

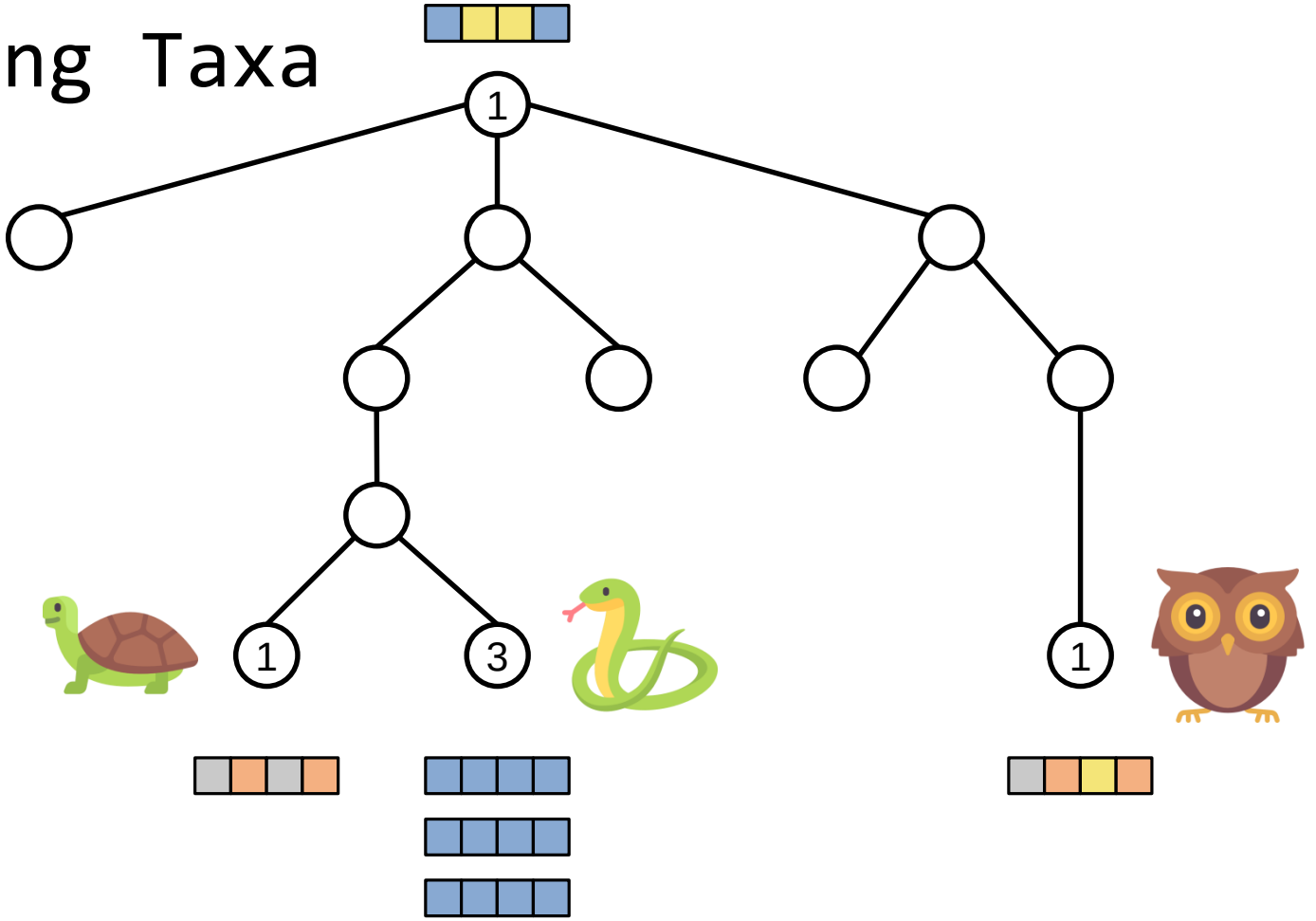
k-mer	observations	classification
T C C T	🦉 🐢	“vertebrate”
T T T T	🐍	“snake”
A G C G	🦉	“owl”
G G C C	🐢 🐍	“reptile”
A G A G	🐢 🐍	“reptile”
C T G A	🦉	“owl”
T A A T	🦉 🐢 🐍	“vertebrate”

Assigning Taxa

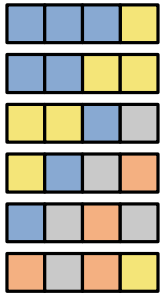


"From what organism did this read originate?"

Assigning Taxa

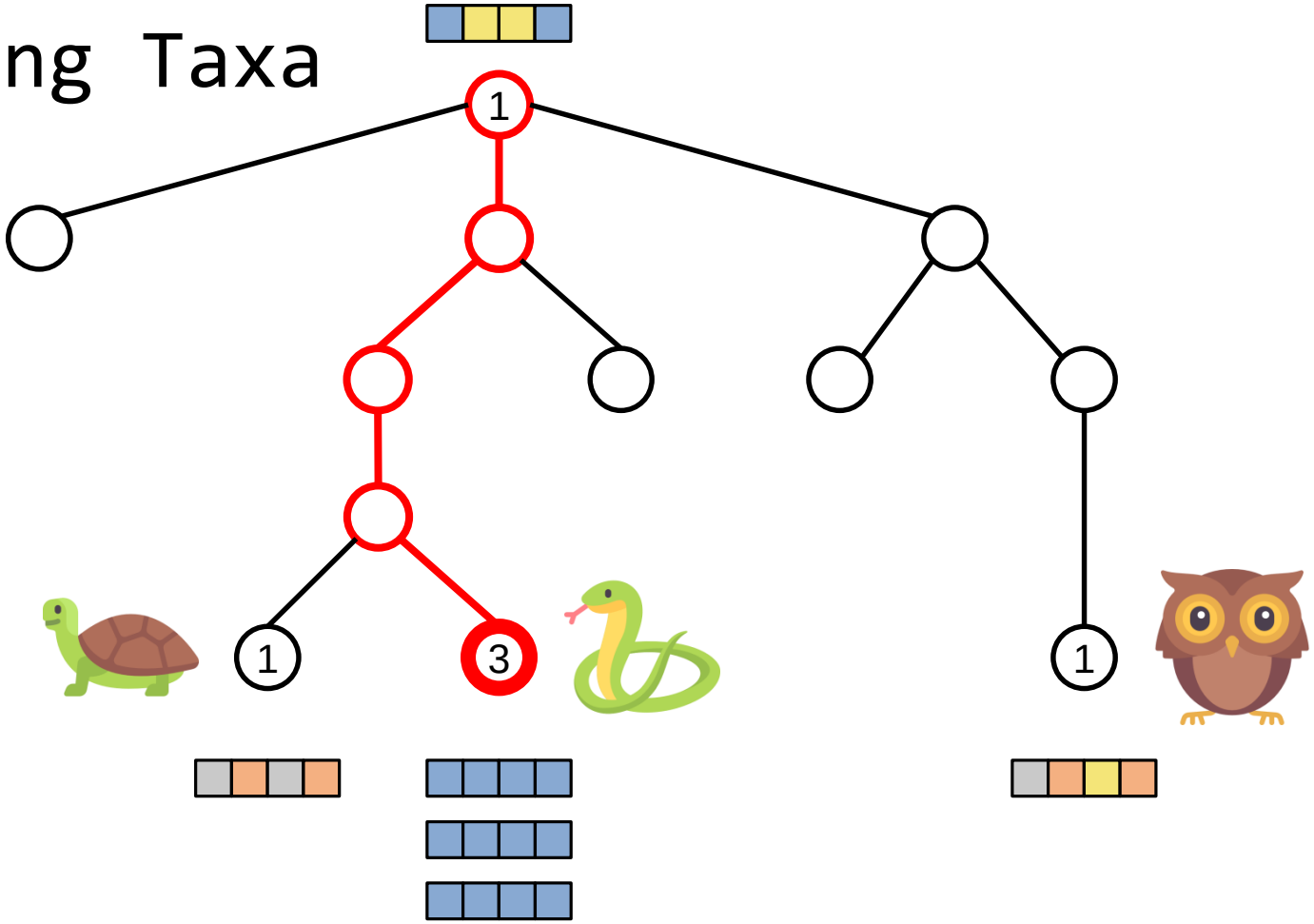


unclassified

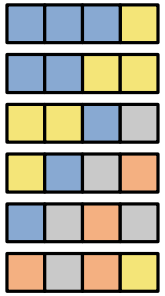


“From what organism did this read originate?”

Assigning Taxa

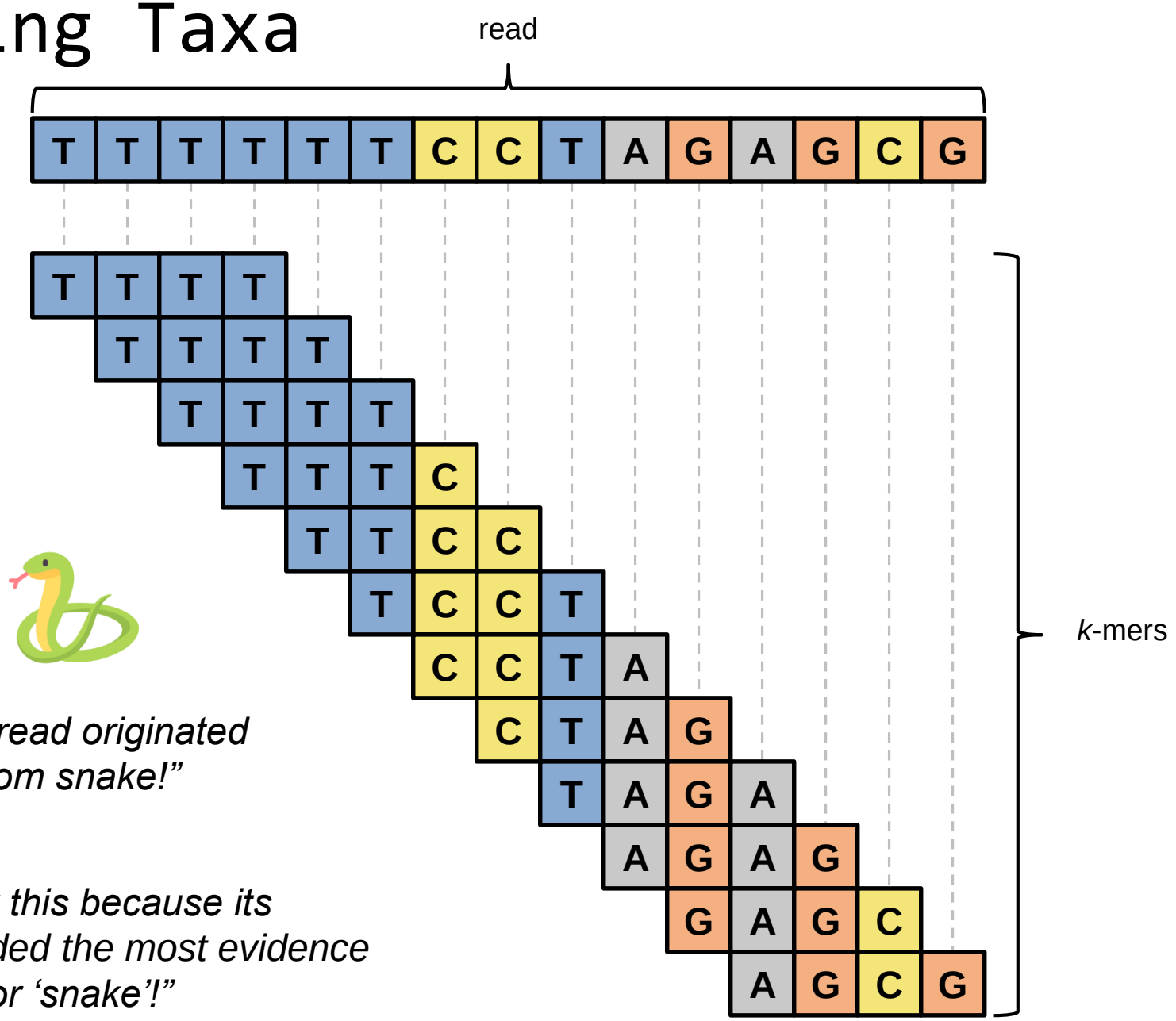


unclassified



The read is classified according to the “root-to-leaf” path with the maximal weight.

Assigning Taxa

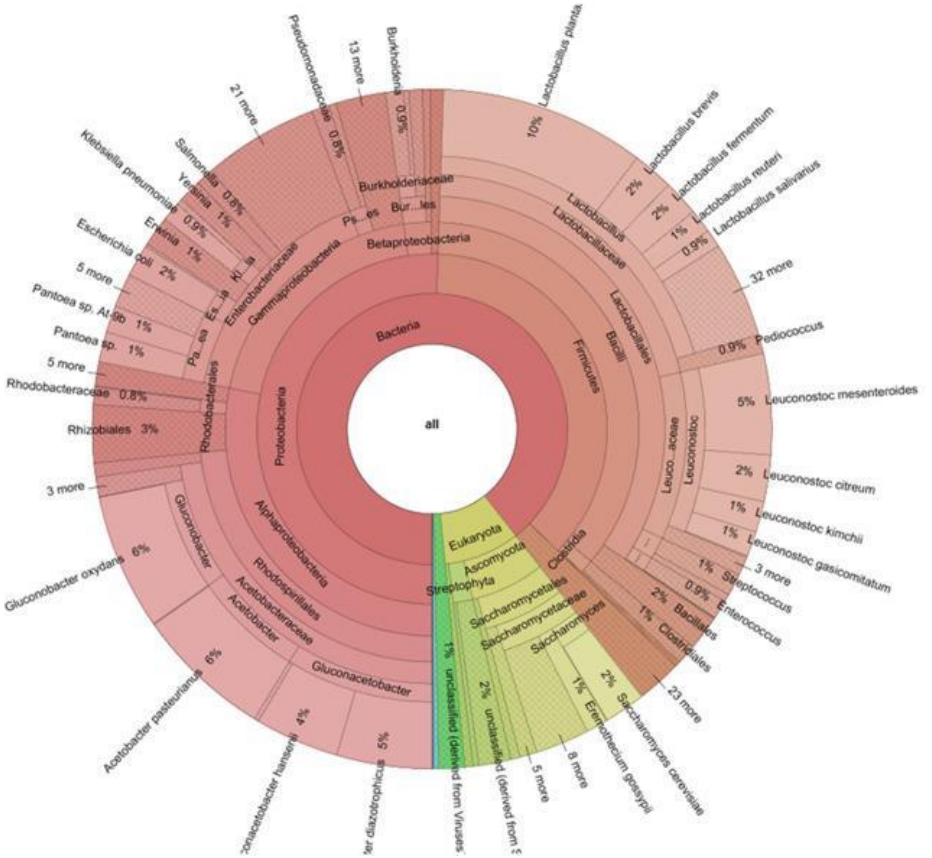


“This read originated from snake!”

“I know this because its k-mers provided the most evidence for ‘snake’!”

Organization, Visualization, and Interpretation

- Krona is a program for visualizing the relative abundances of taxa in a web browser.



Ondov BD, et al. "Interactive metagenomic visualization in a Web browser". 2011. https://doi.org/10.1186/1471-2105-12-385

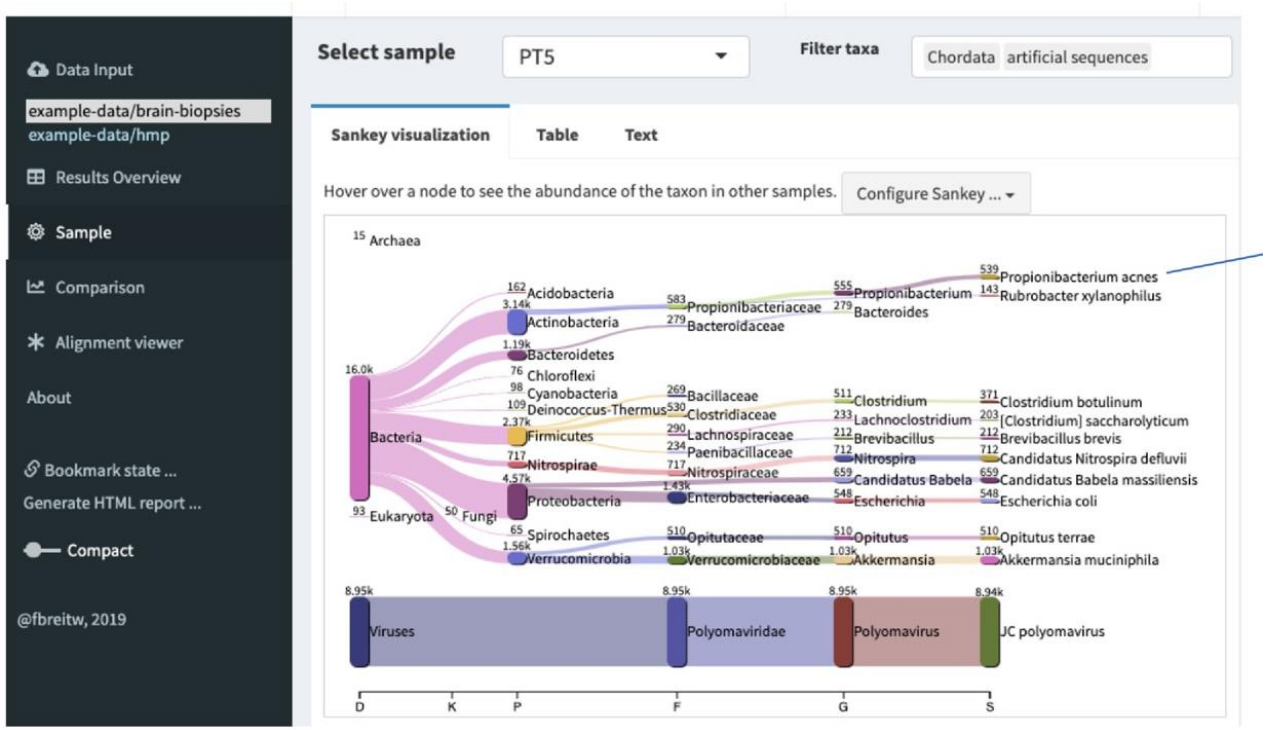
Organization, Visualization, and Interpretation

- Krona is a program for visualizing the relative abundances of taxa in a web browser.
- Pavian is a newer package that supports comparisons of taxa from multiple datasets at once.

Breitwieser FP, Salzberg SL. "Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification". 2019. https://doi.org/10.1093/bioinformatics/btz715

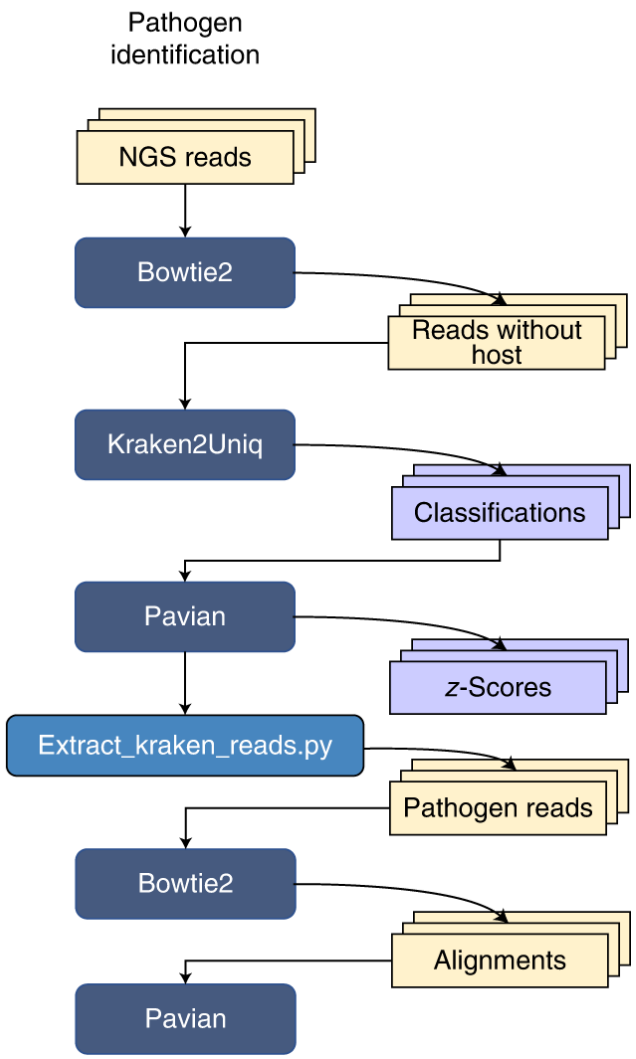
Organization, Visualization, and Interpretation

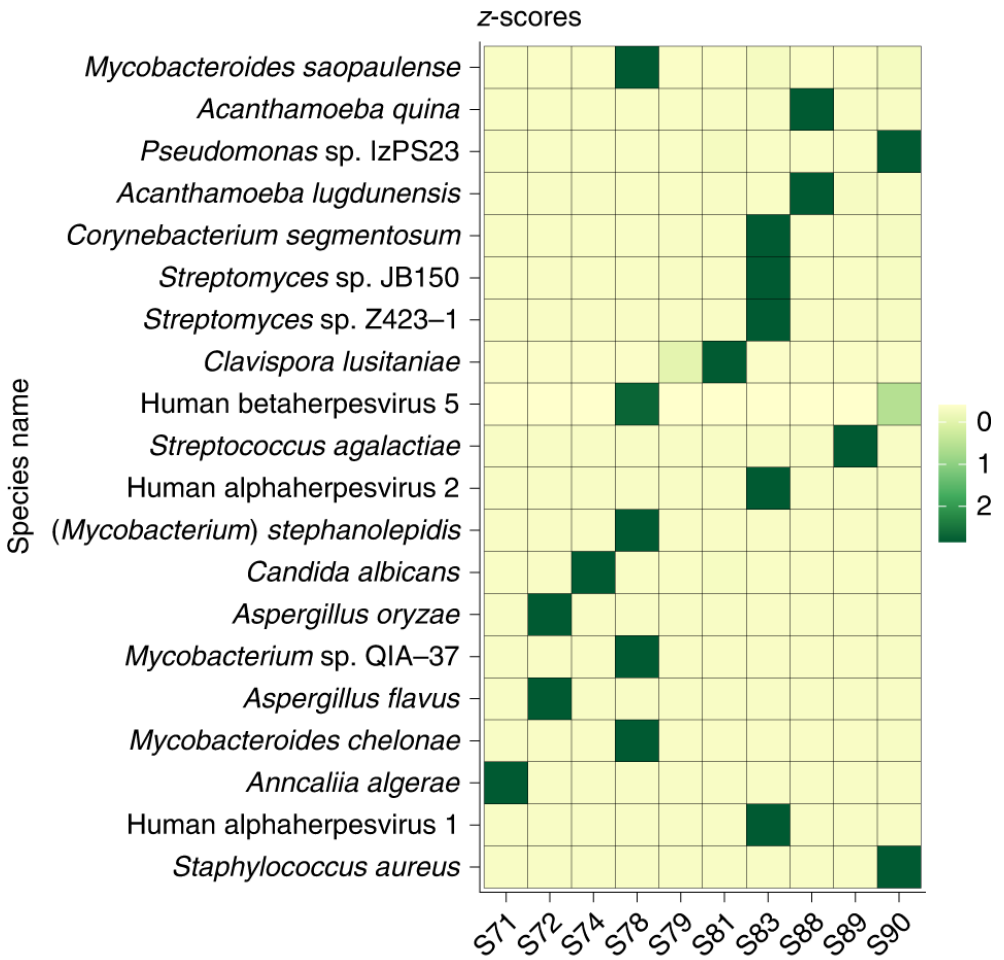
- Krona is a program for visualizing the relative abundances of taxa in a web browser.
- Pavian is a newer package that supports comparisons of taxa from multiple datasets at once.
- How do you know which taxa are the causative agent of the disease?
  - Compare the symptoms of the pathogen with the symptoms of the disease.
  - Comparison to controls.



Breitwieser FP, Salzberg SL. "Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification". 2019. https://doi.org/10.1093/bioinformatics/btz715

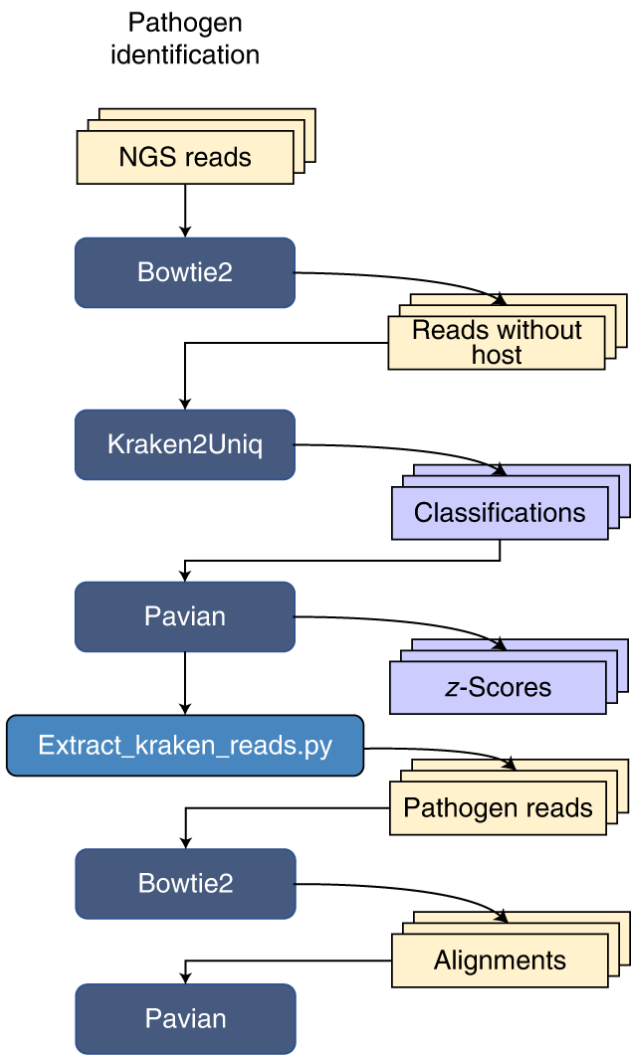
Metagenomics analysis protocol using Kraken

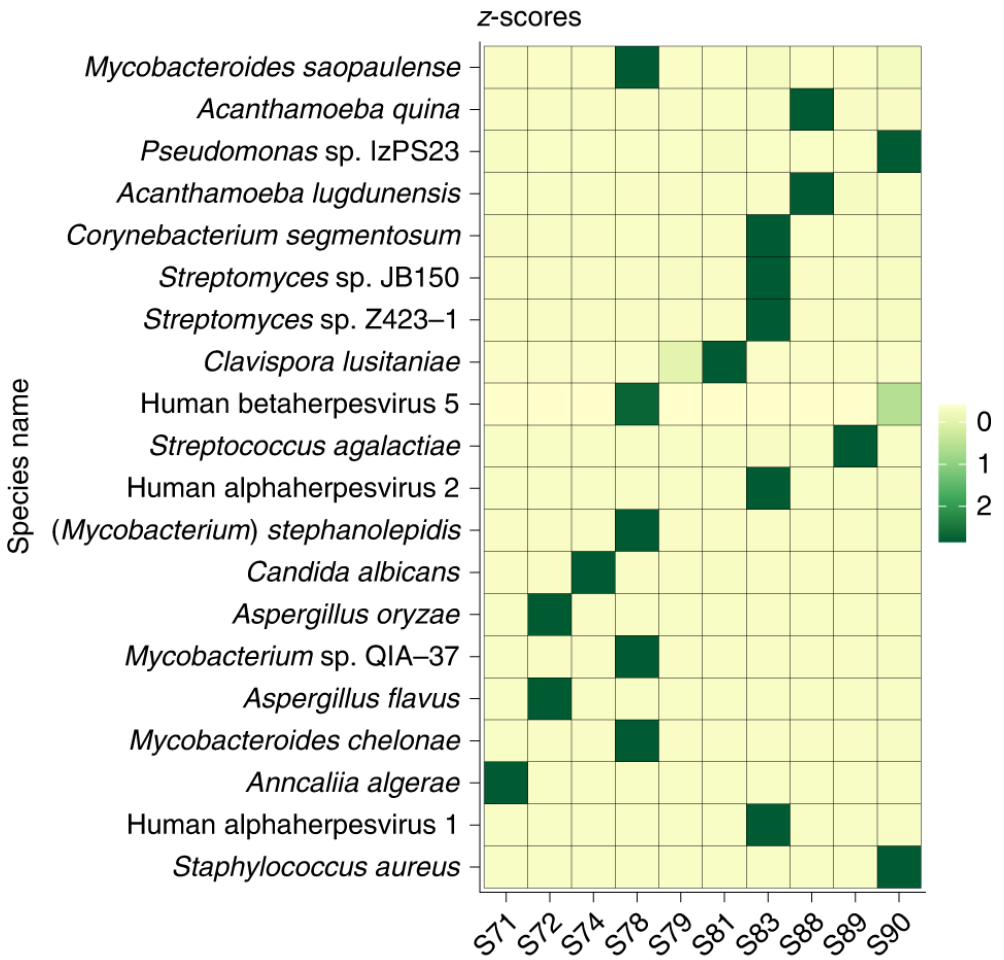




Lu J, et al. "Metagenome analysis using the Kraken software suite". 2022. https://doi.org/10.1038/s41596-022-00738-y

Metagenomics analysis protocol using Kraken





For our lab we will use different data and a simplified version of this pathogen identification protocol with additional steps to include metagenomics assembly.

Lu J, et al. "Metagenome analysis using the Kraken software suite". 2022. https://doi.org/10.1038/s41596-022-00738-y

We are on a Coffee Break & Networking Session

Workshop Sponsors:

Canadian Centre for Computational Genomics