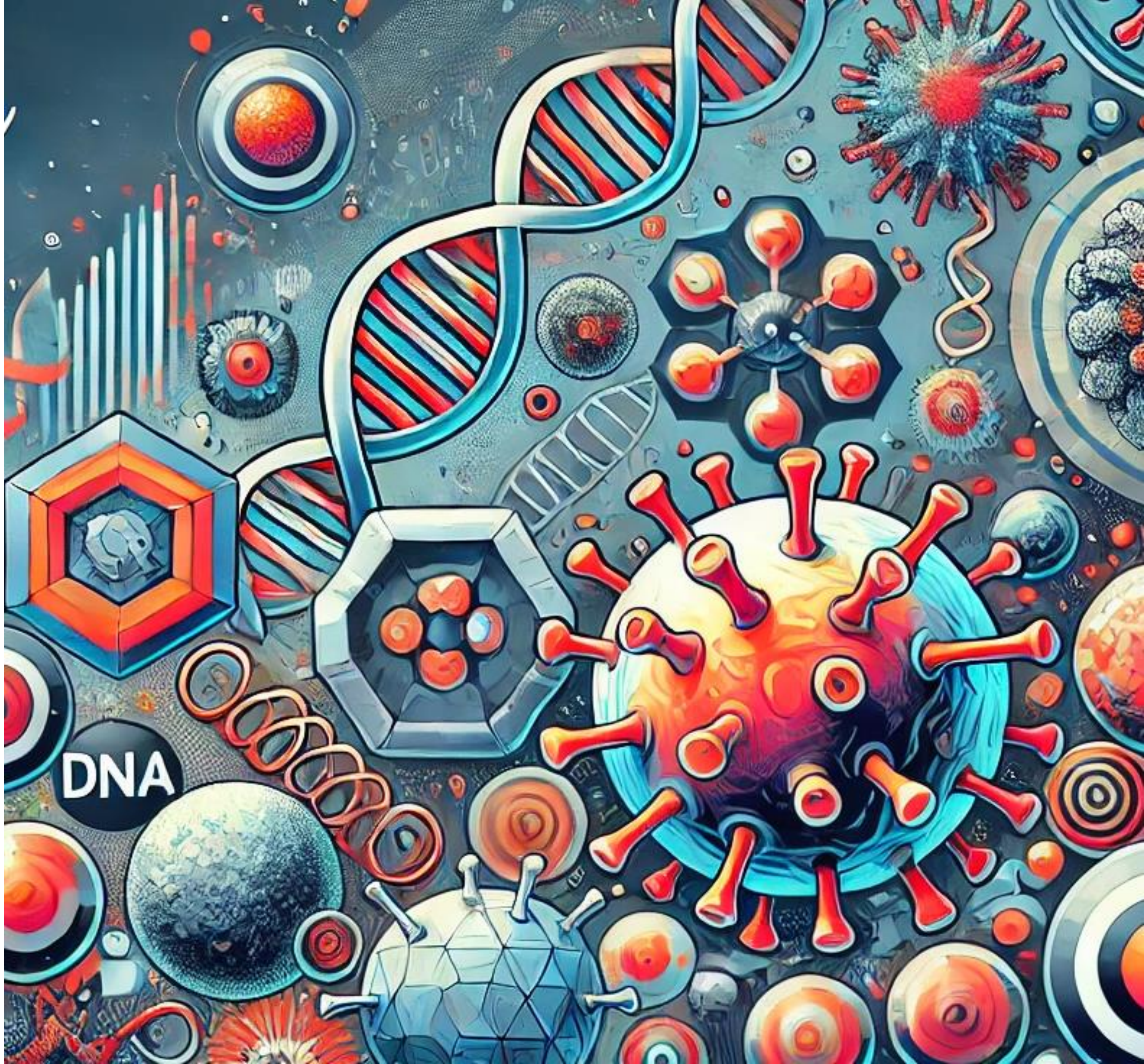
July 30, 2024

Pathogen Discovery, Identification and Characterization in the Era of Metagenomics

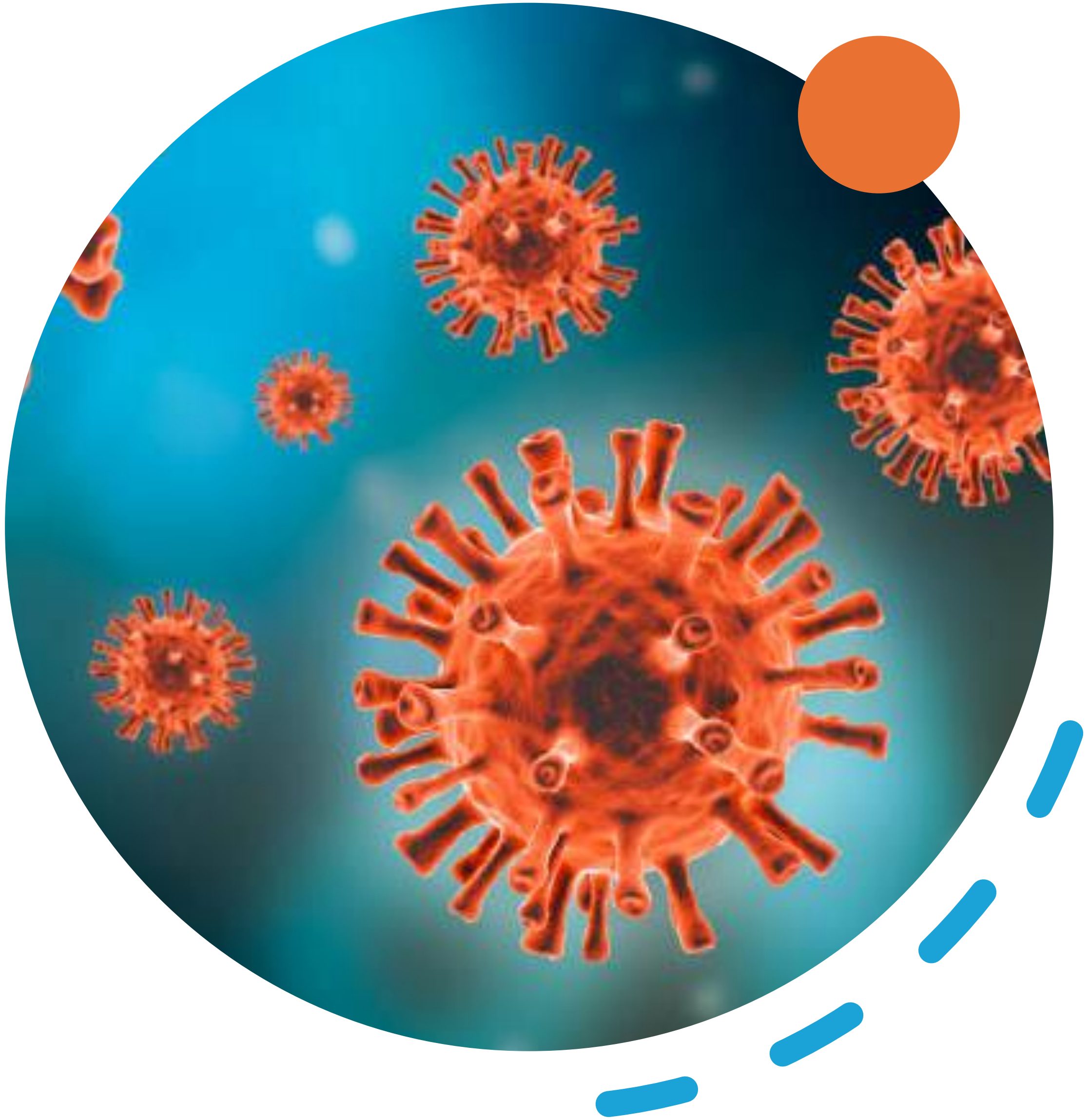
Arunee Thitithanyanont, M.D.

Department of Microbiology
Faculty of Science, Mahidol University

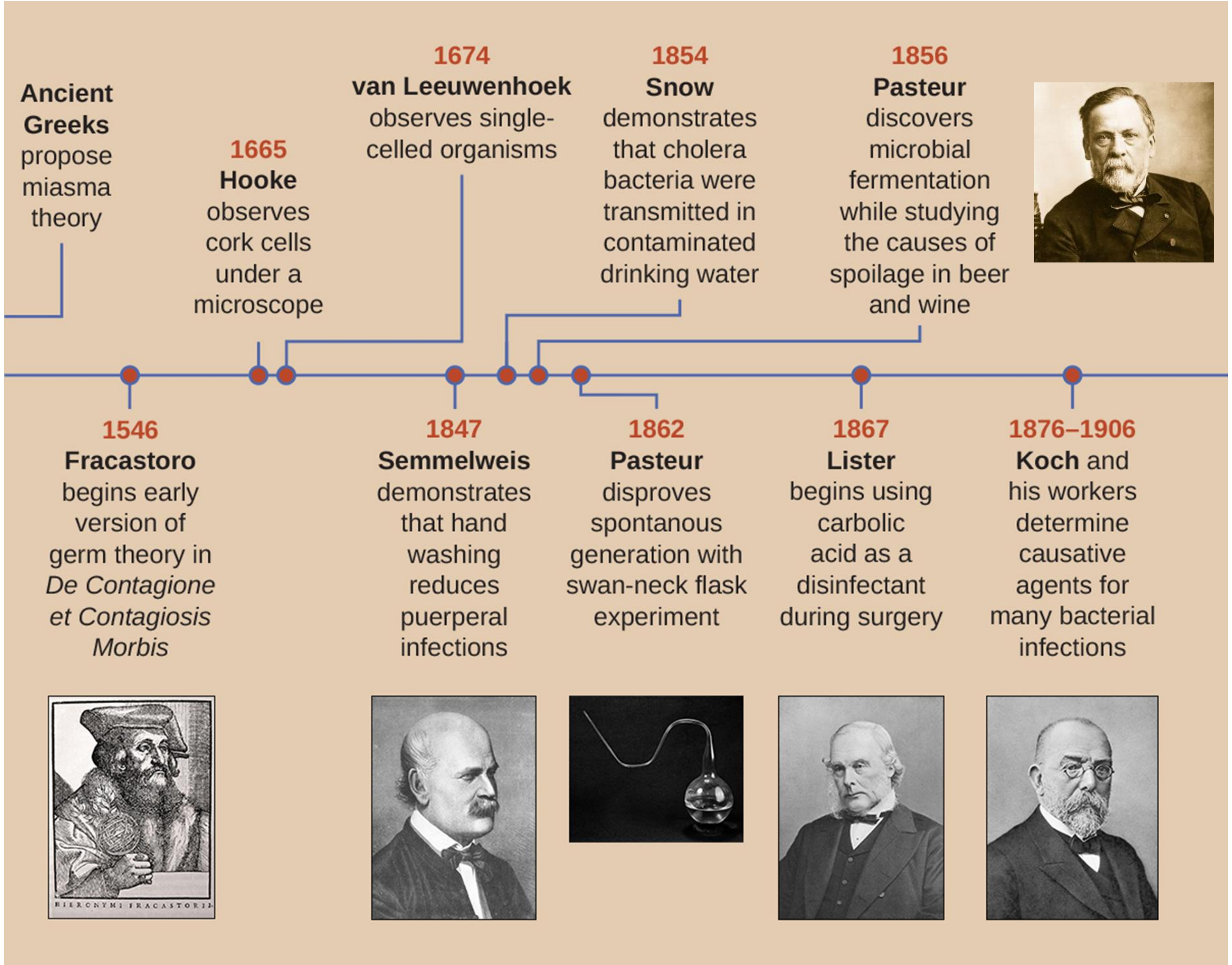


Outline

- Introduction
- Traditional vs. Metagenomic Approaches
- Case Study: Pathogen Discovery
- Challenges in Metagenomics
- Clinical Applications
- Future Directions



The Germ Theory of Diseases



Ancient Greeks propose miasma theory

1546
Fracastoro begins early version of germ theory in *De Contagione et Contagiosis Morbis*

1665
Hooke observes cork cells under a microscope

1674
van Leeuwenhoek observes single-celled organisms

1847
Semmelweis demonstrates that hand washing reduces puerperal infections

1854
Snow demonstrates that cholera bacteria were transmitted in contaminated drinking water

1856
Pasteur discovers microbial fermentation while studying the causes of spoilage in beer and wine

1862
Pasteur disproves spontanous generation with swan-neck flask experiment

1867
Lister begins using carbolic acid as a disinfectant during surgery

1876–1906
Koch and his workers determine causative agents for many bacterial infections

Microorganisms cause diseases.

Limitation of Koch’s Postulates

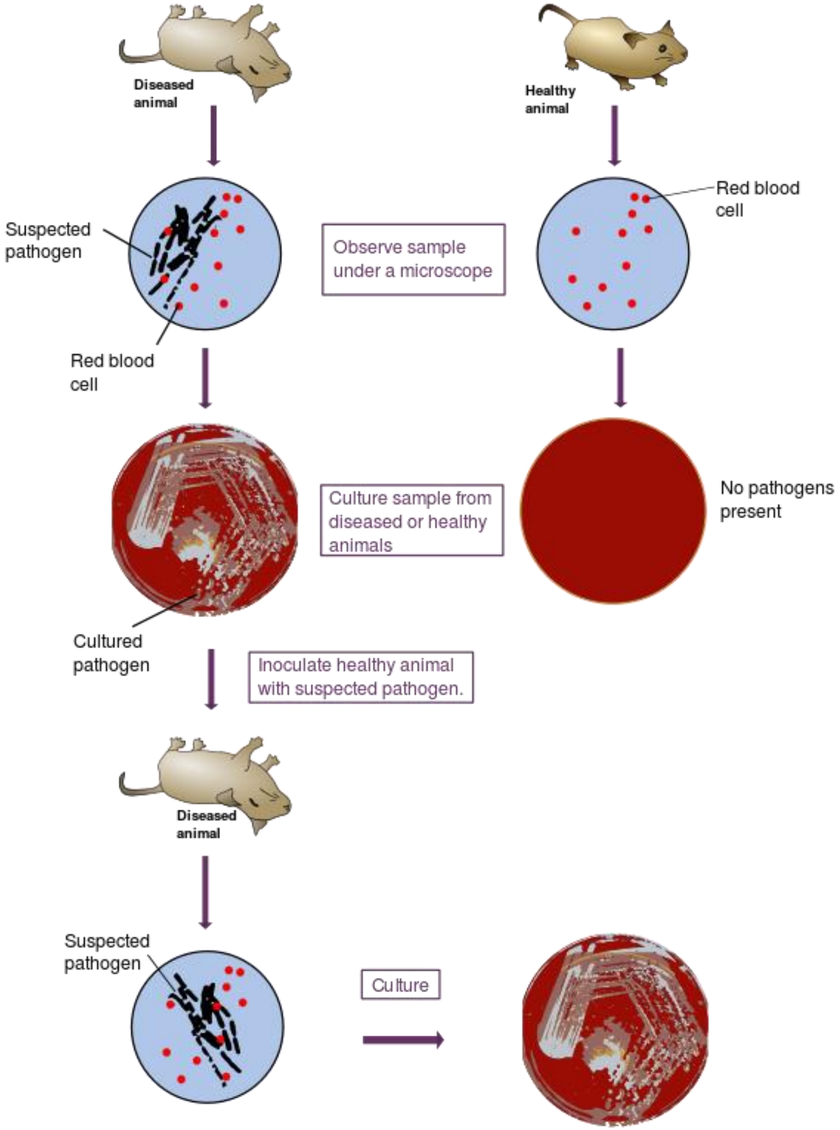
Koch's Postulates:

① The microorganism must be found in abundance in all organisms suffering from the disease, but should not be found in healthy organisms.

② The microorganism must be isolated from a diseased organism and grown in pure culture.

③ The cultured microorganism should cause disease when introduced into a healthy organism.

④ The microorganism must be reisolated from the inoculated, diseased experimental host and identified as being identical to the original specific causative agent.



Koch’s postulates is sufficient but not necessary to establish causation

Asymptomatic Carriers

Inability to Culture

Ethical Concerns

Complex Interactions

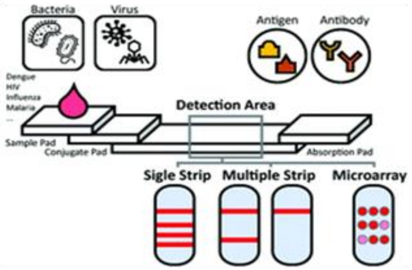
Koch’s Postulates cannot be universally applied, especially to viral diseases.

Infectious diseases diagnostic approaches

Culture based

Traditional methods

Serology based

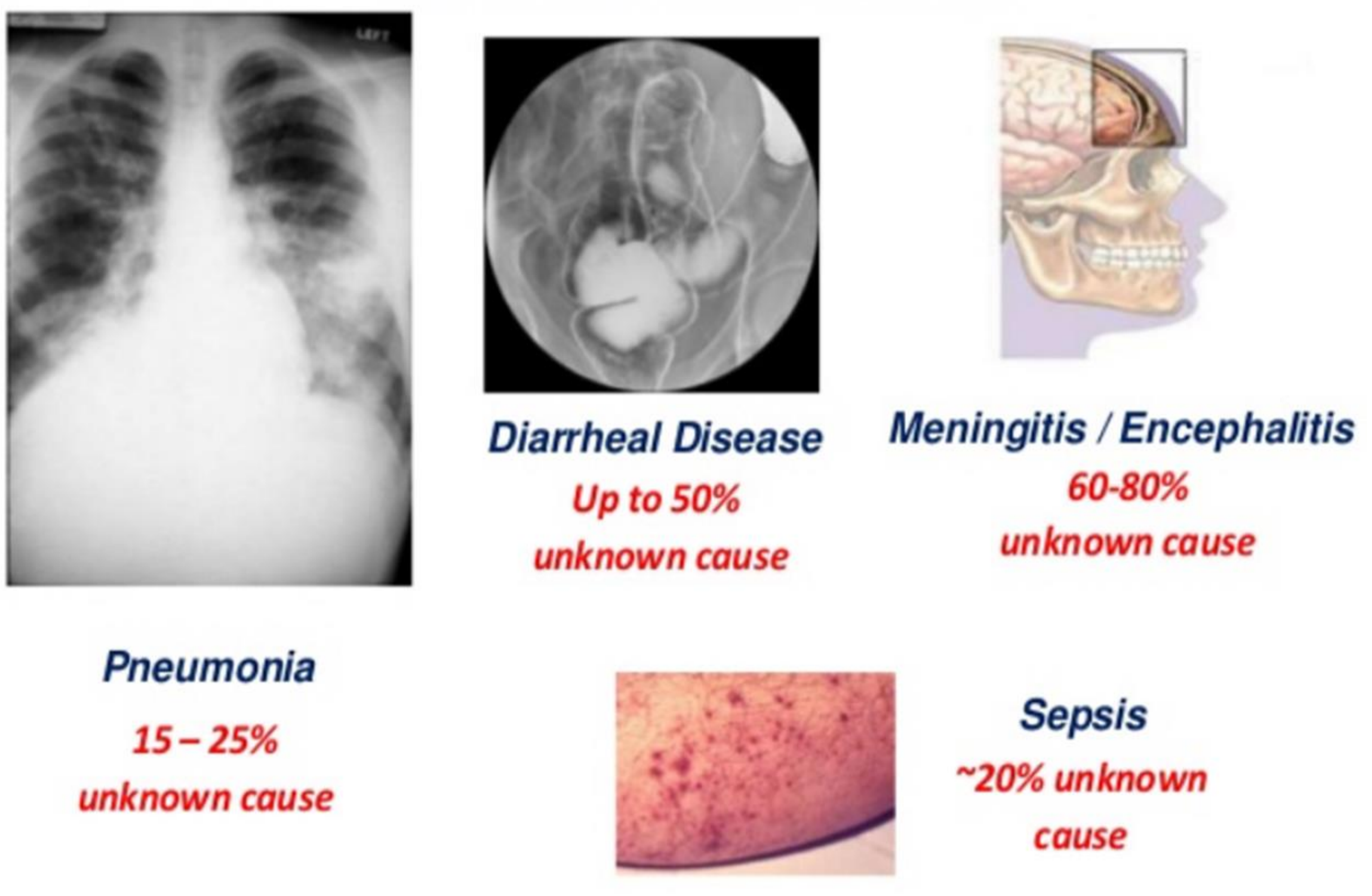


Molecular methods

NGS based

PCR based

Laboratory Failure to Identify Pathogens: Inadequacy of Current Tools



Culture

- Labor intensive
- Need for special media
- Prolonged period of time to culture
- Some organisms are uncultivable on artificial media
- Potential health hazards

Antigen Detection

- Negative tests require confirmation
- Effected by poor specimen collection
- Low microbe burden

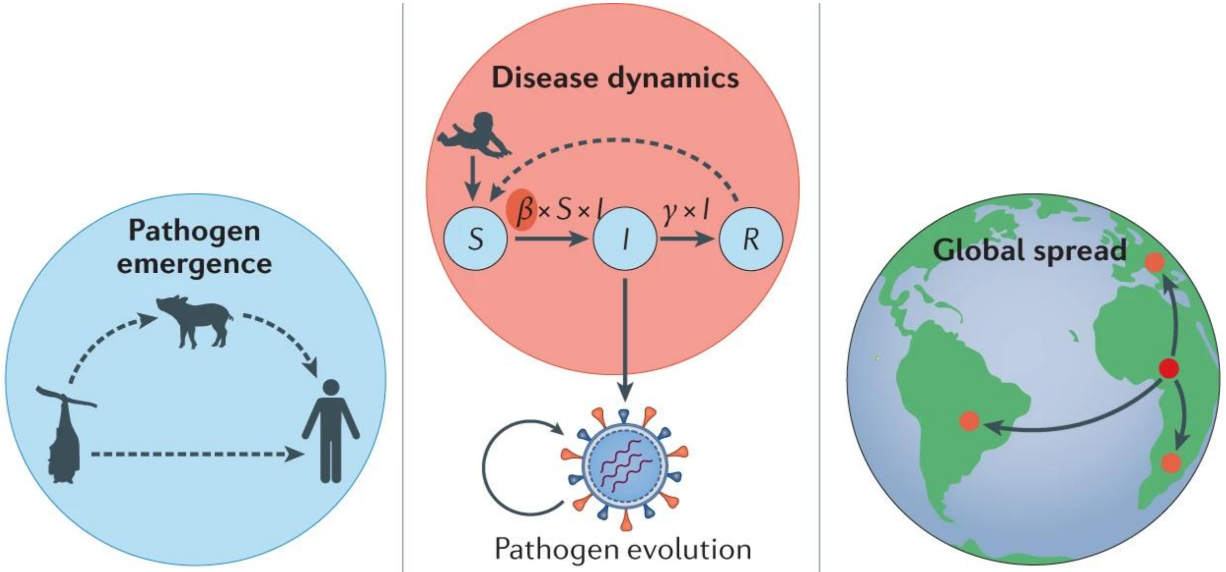
Serology

- Unhelpful during early stage of infection

Chiu et al. Nature Reviews Genetics volume 20, (2019)

A changing world requires changing science

- Pathogen emergence
- Disease dynamics
- Global spread



Climatic change	Drives range shifts for reservoir species	Affects transmission and susceptibility	Affects the geographical range of vectors
Technological change			
Transportation	Improved global surveillance		Air transit and high-speed rail affect pace and range of spread
Health care		Vaccination affects dynamics	Improved care reduces burden
Demographic change			
Population growth and land use	Increased contact with reservoir species	Population numbers affect evolution, birth rates affect dynamics	Larger population travelling
Urbanization	Depends on species	Density affects contact rate	Urban population more connected
Ageing	Imunosenescence affects spillover risk	Ageing population increases transmission	Possible larger burden

Nat Rev Microbiol (2021)

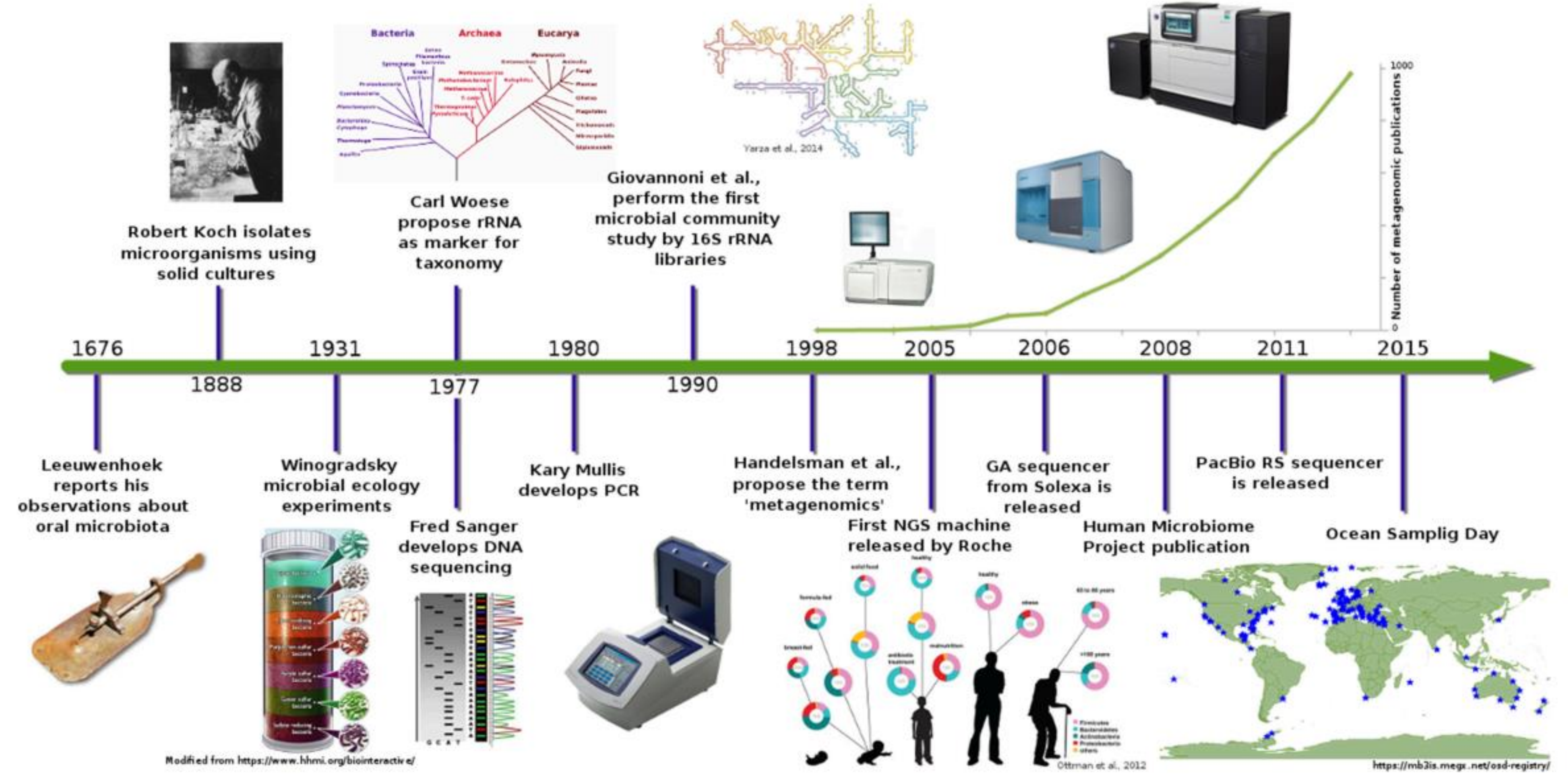
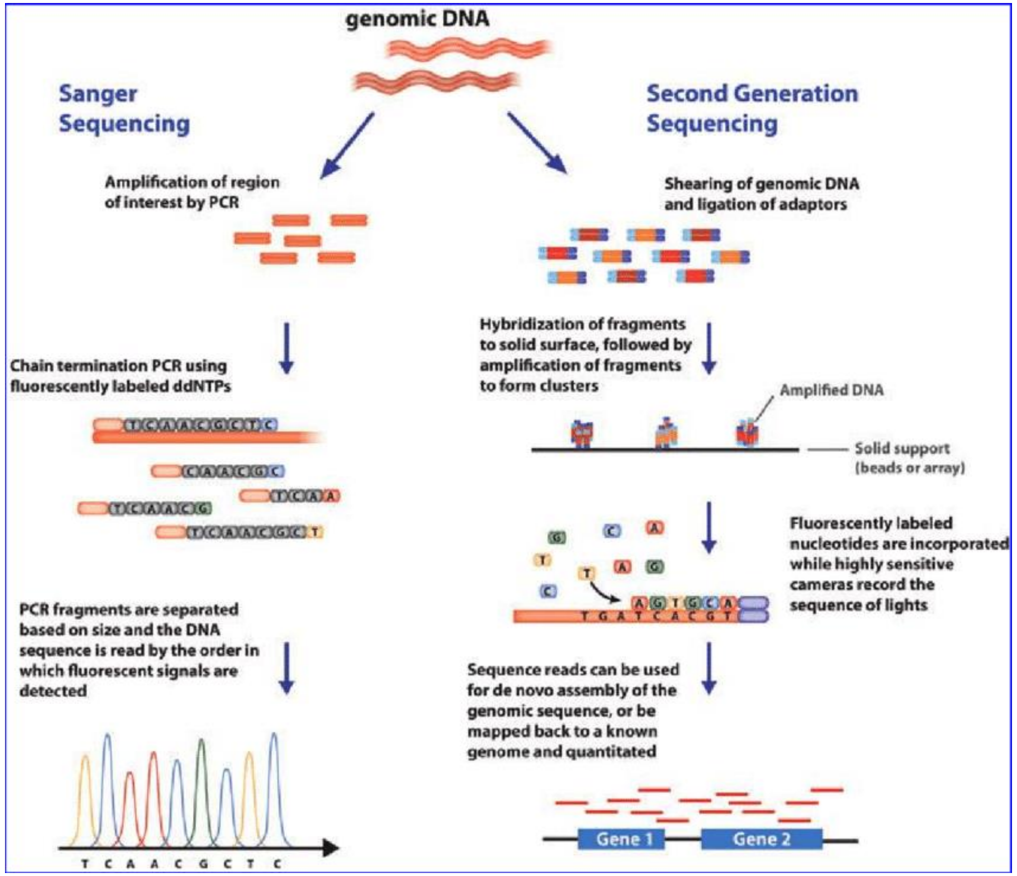
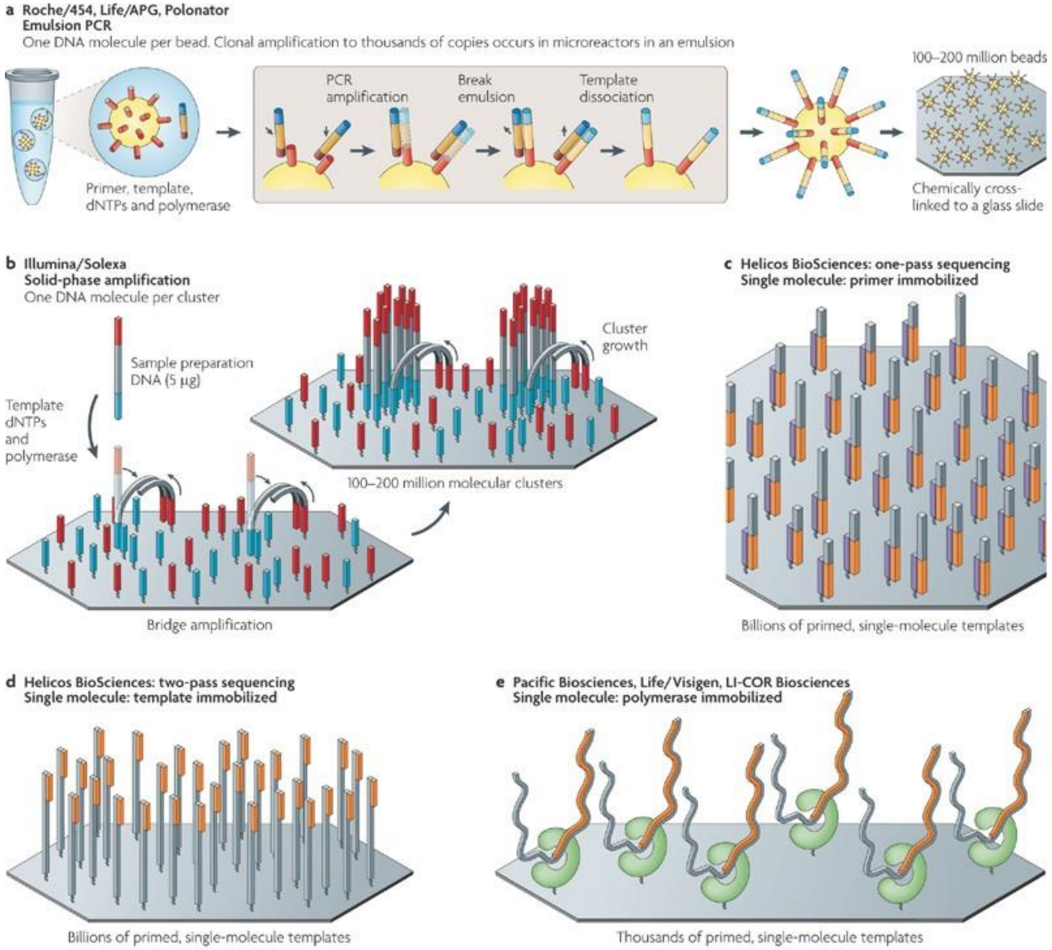


FIGURE 1 | Metagenomics timeline and milestones. Timeline showing advances in microbial communities studies from Leeuwenhoek to NGS (Ottman et al., 2012; Yarza et al., 2014).

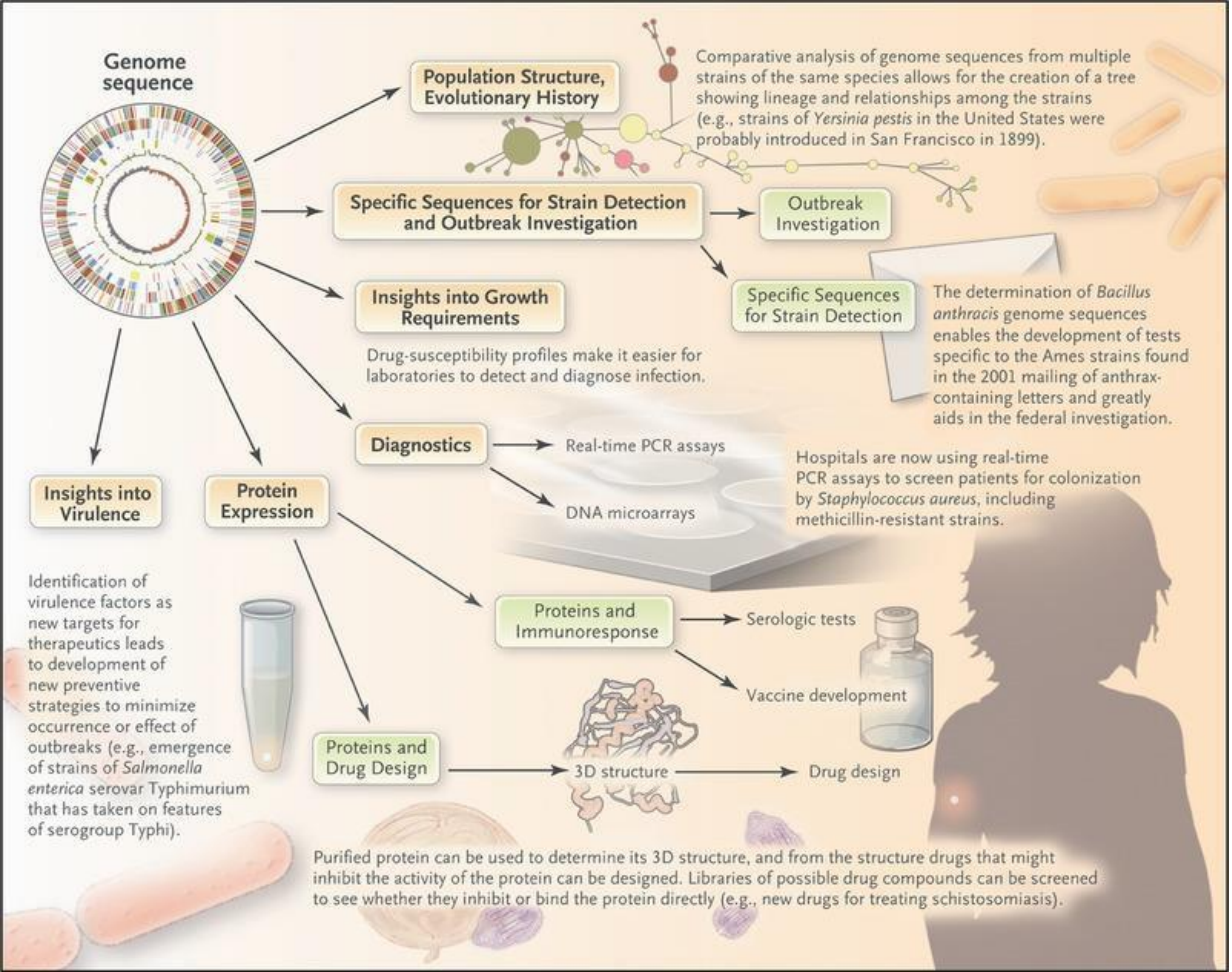
Next generation sequencing (NGS)





Nature Reviews | Genetics

Microbial Genomics and Tool Development



Comparative analysis of genome sequences from multiple strains of the same species allows for the creation of a tree showing lineage and relationships among the strains (e.g., strains of *Yersinia pestis* in the United States were probably introduced in San Francisco in 1899).

The determination of *Bacillus anthracis* genome sequences enables the development of tests specific to the Ames strains found in the 2001 mailing of anthrax-containing letters and greatly aids in the federal investigation.

Drug-susceptibility profiles make it easier for laboratories to detect and diagnose infection.

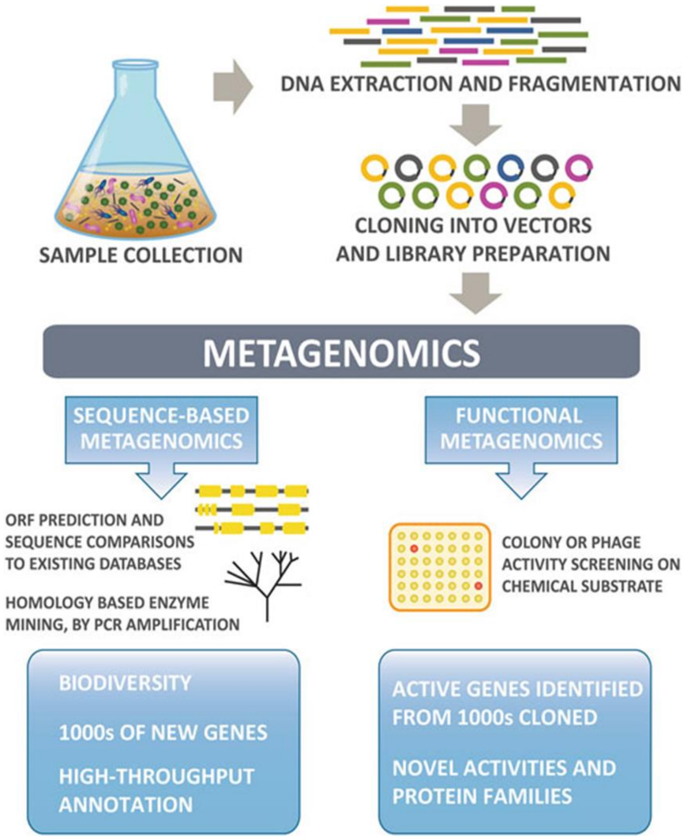
Hospitals are now using real-time PCR assays to screen patients for colonization by *Staphylococcus aureus*, including methicillin-resistant strains.

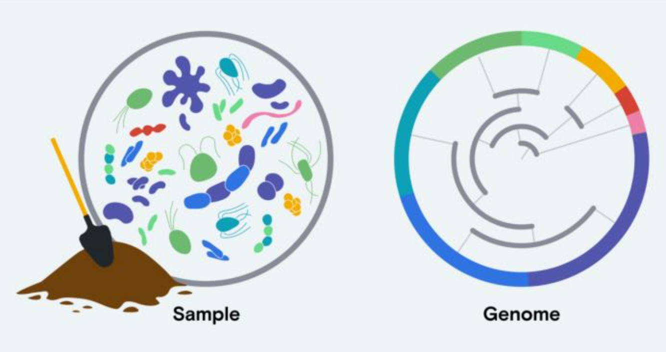
Identification of virulence factors as new targets for therapeutics leads to development of new preventive strategies to minimize occurrence or effect of outbreaks (e.g., emergence of strains of *Salmonella enterica* serovar Typhimurium that has taken on features of serogroup Typhi).

Purified protein can be used to determine its 3D structure, and from the structure drugs that might inhibit the activity of the protein can be designed. Libraries of possible drug compounds can be screened to see whether they inhibit or bind the protein directly (e.g., new drugs for treating schistosomiasis).

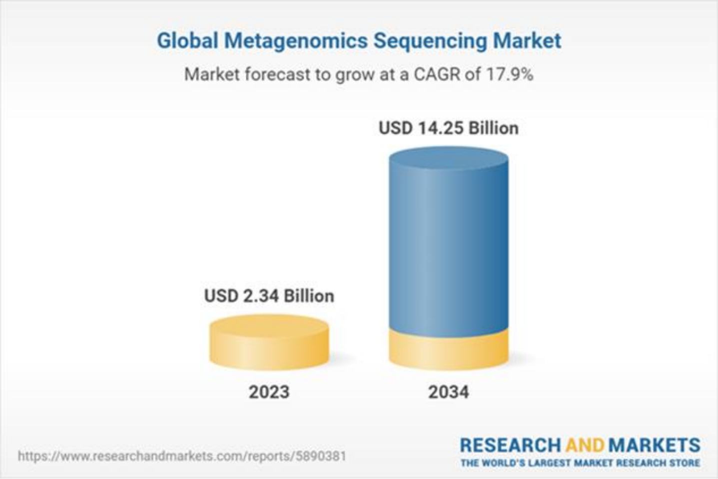
N Engl J Med 2011; 365:347-357
DOI: 10.1056/NEJMra1003071

Metagenomics allows pan-nucleic acid detection directly from specimens





Metagenomics studies the structure and function of nucleotide sequences from all organisms in a sample.

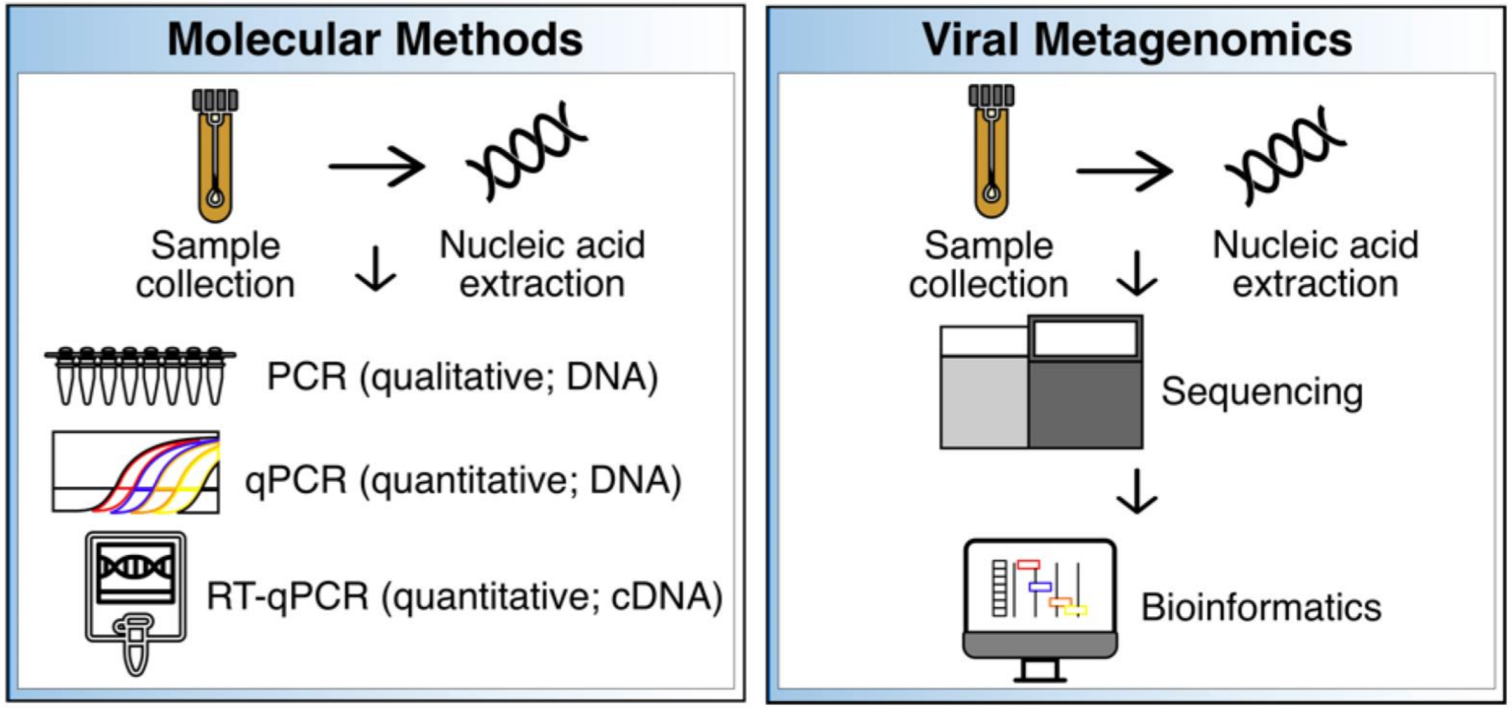

Global Metagenomics Sequencing Market

Market forecast to grow at a CAGR of 17.9%

https://www.researchandmarkets.com/reports/5890381

https://www.genome.gov/genetics-glossary/Metagenomics

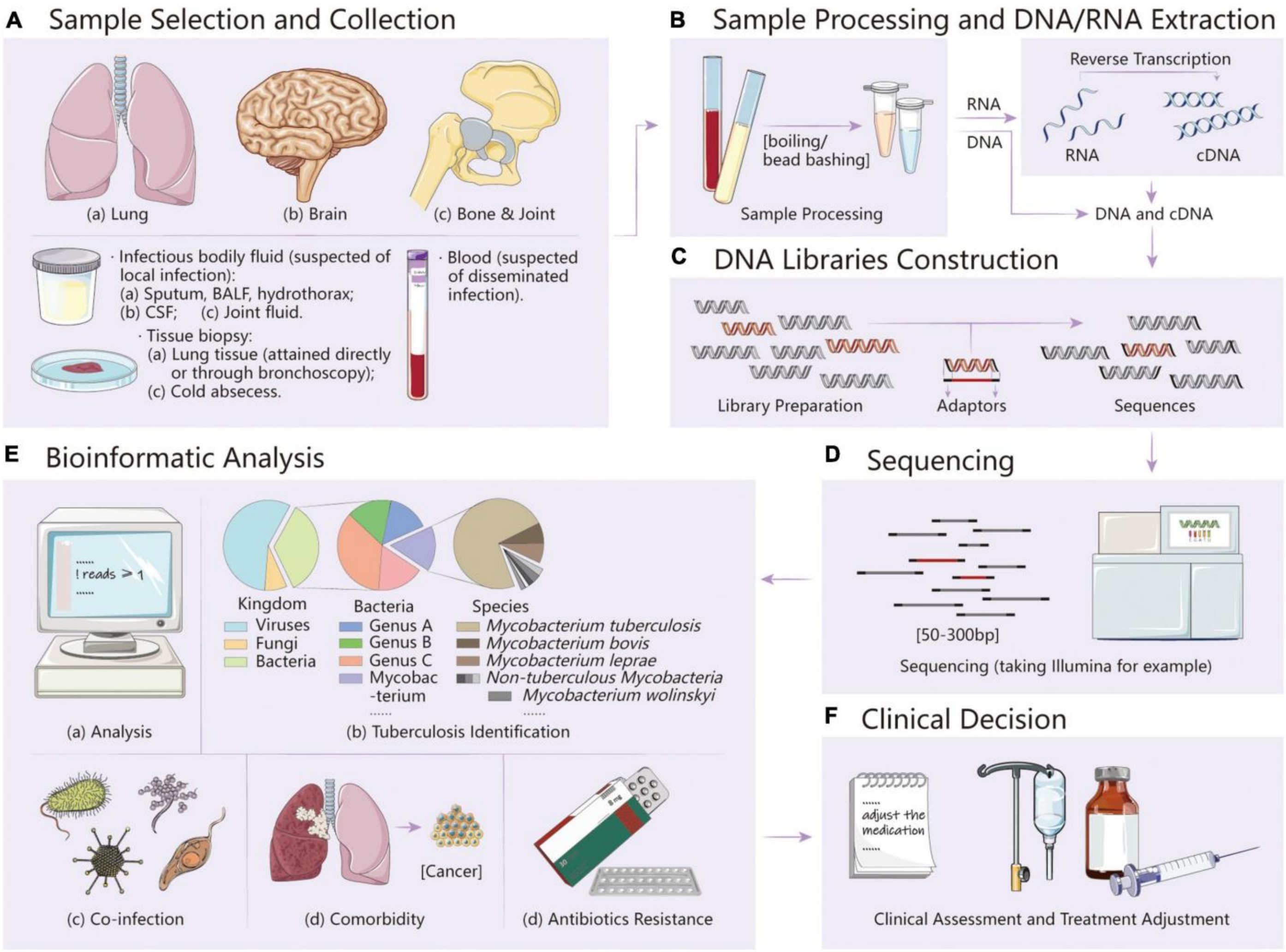
Molecular method vs Metagenomics



Molecular Method:
- Analyzes specific genes from cultured organisms.
- Requires prior knowledge of the target.

Metagenomics:
- Analyzes all genetic material from environmental samples.
- No culturing or prior knowledge

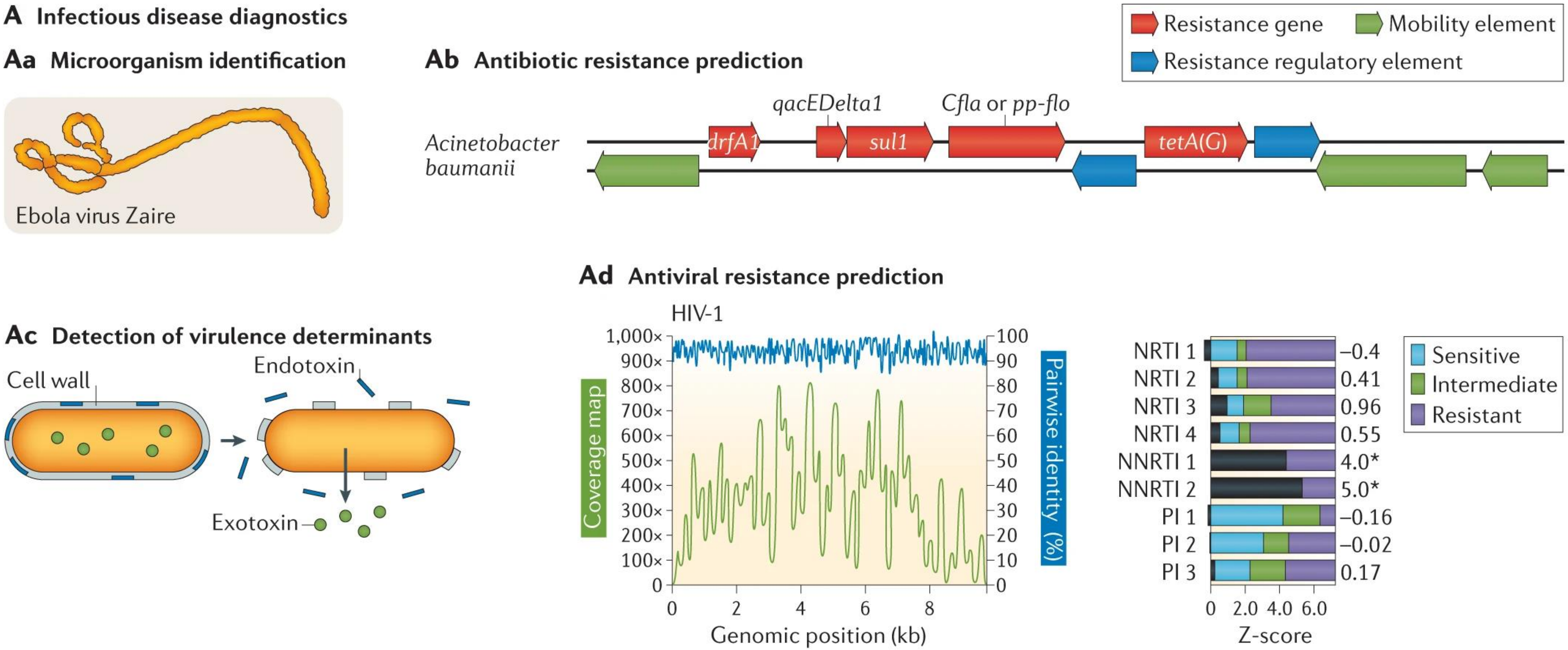
A typical workflow of mNGS



- Viral phylogeny and taxonomy
- Proviral metagenome
- Viral community structure and ecology
- Snapshot of microbiome
- Investigation of possible co-infection

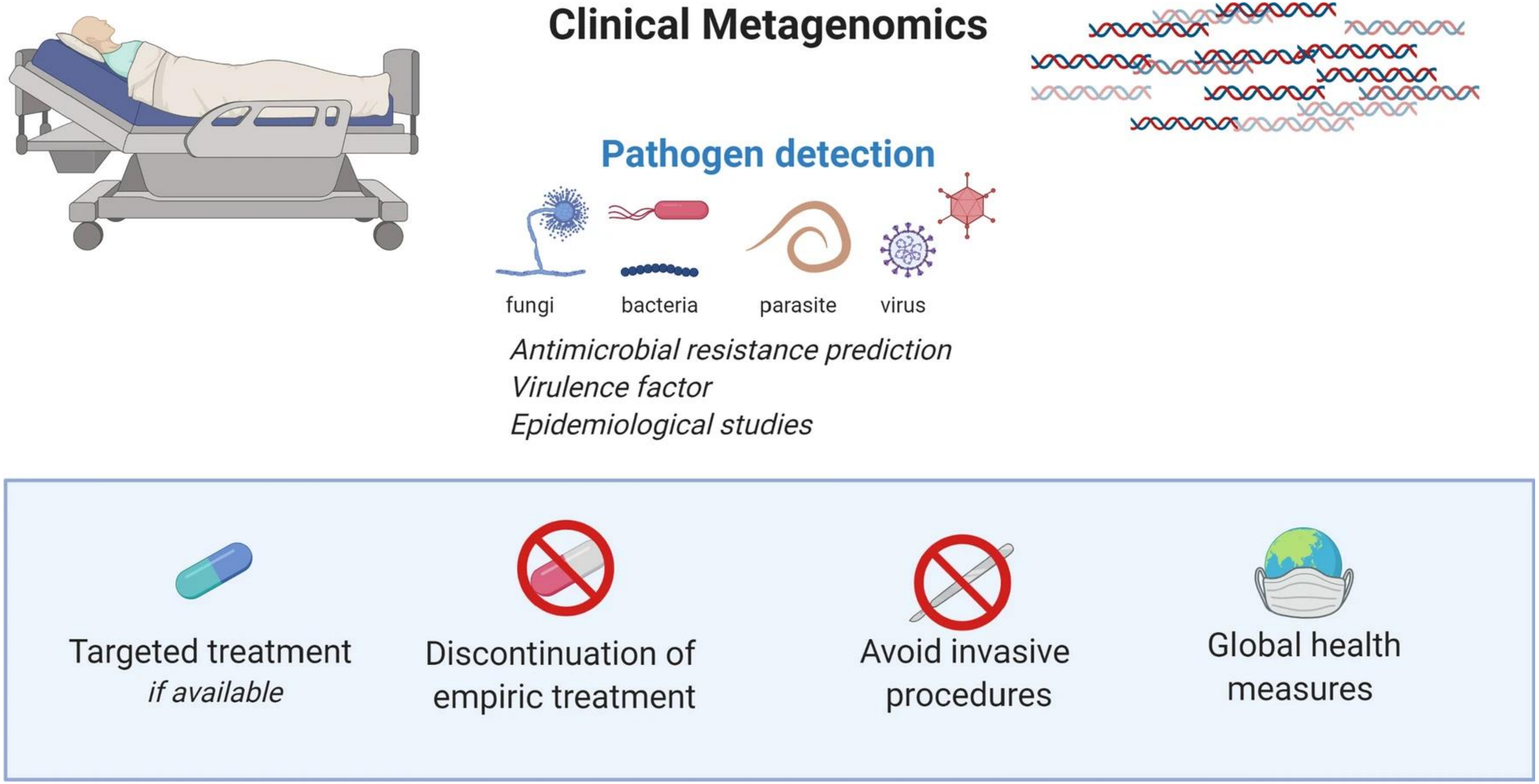
When to use the “Metagenomics” approach”?

Clinical applications of metagenomic sequencing in infectious disease diagnosis



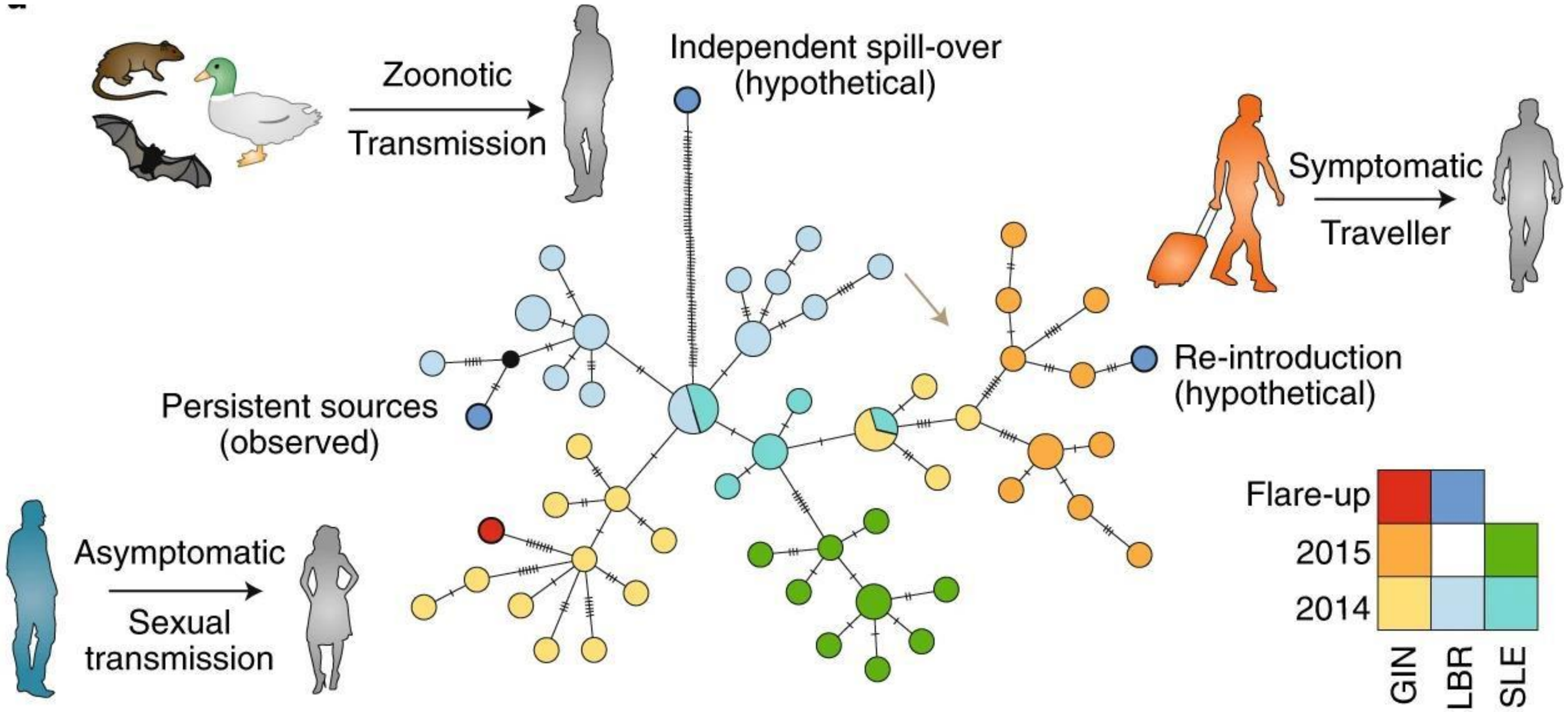
Nature Reviews Genetics **volume 20**, pages341–355 (2019)

Therapeutic Perspectives



Drugs 81, 1453–1466 (2021)

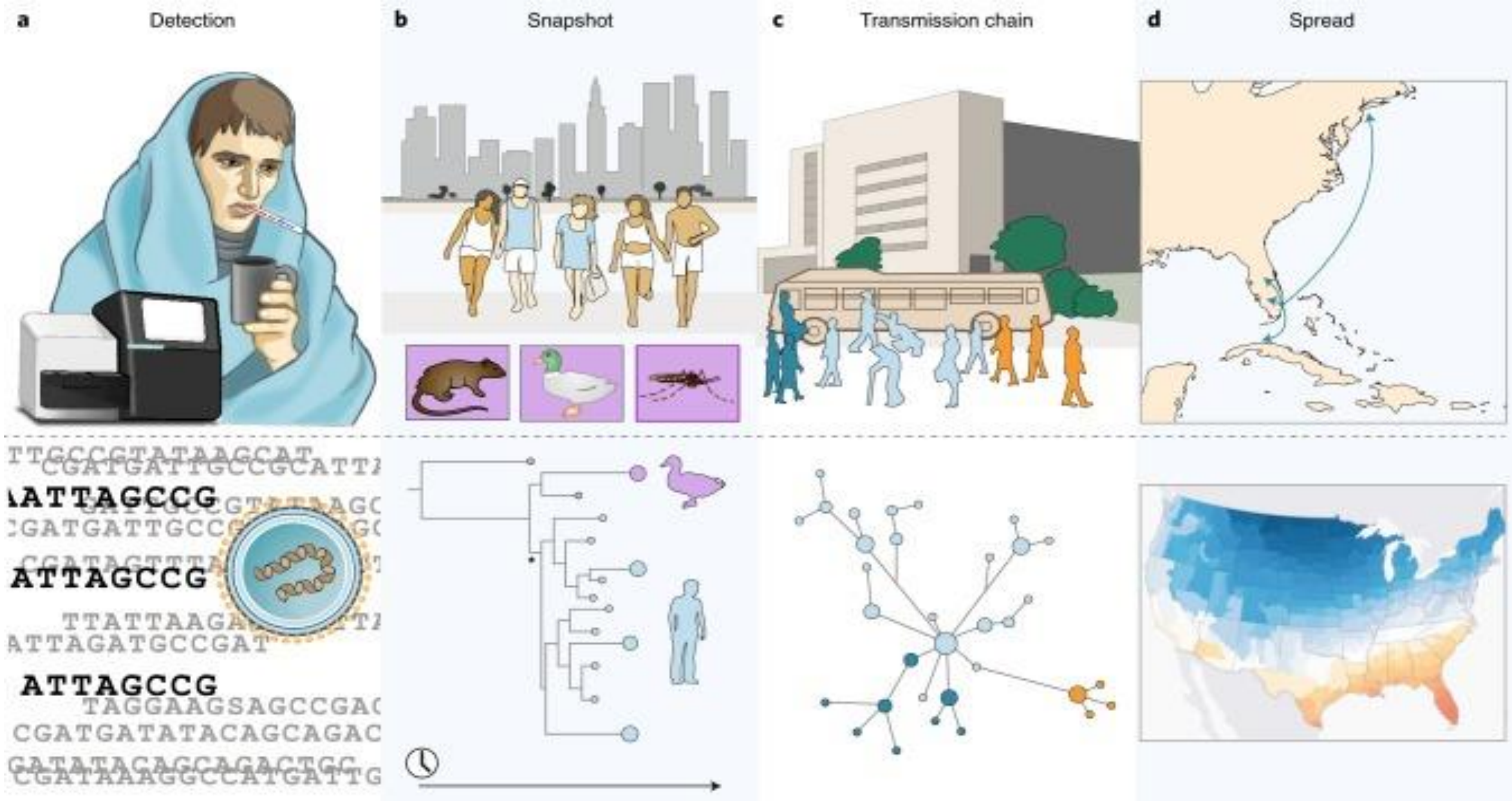
Transmission chain tracking during outbreaks using virus genomics



Nature Microbiology volume 4, pages10–19 (2019)

Outbreak detection:

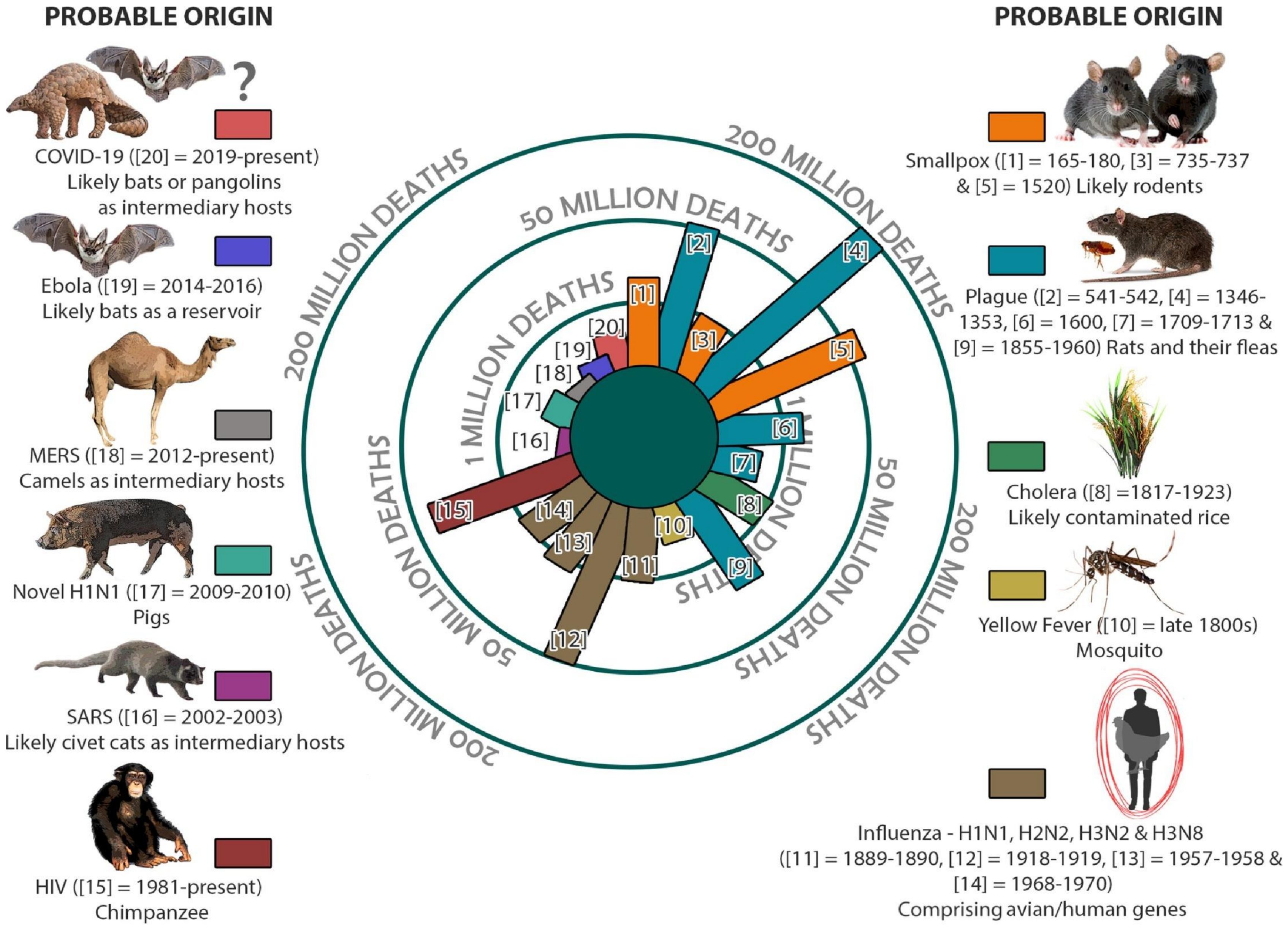
Real-time genomic investigation of **Disease X**



- SARS
- Pandemic influenza
- MERS
- Ebola
- Zika
- COVID19

Nature Microbiology volume 4, pages10–19 (2019)

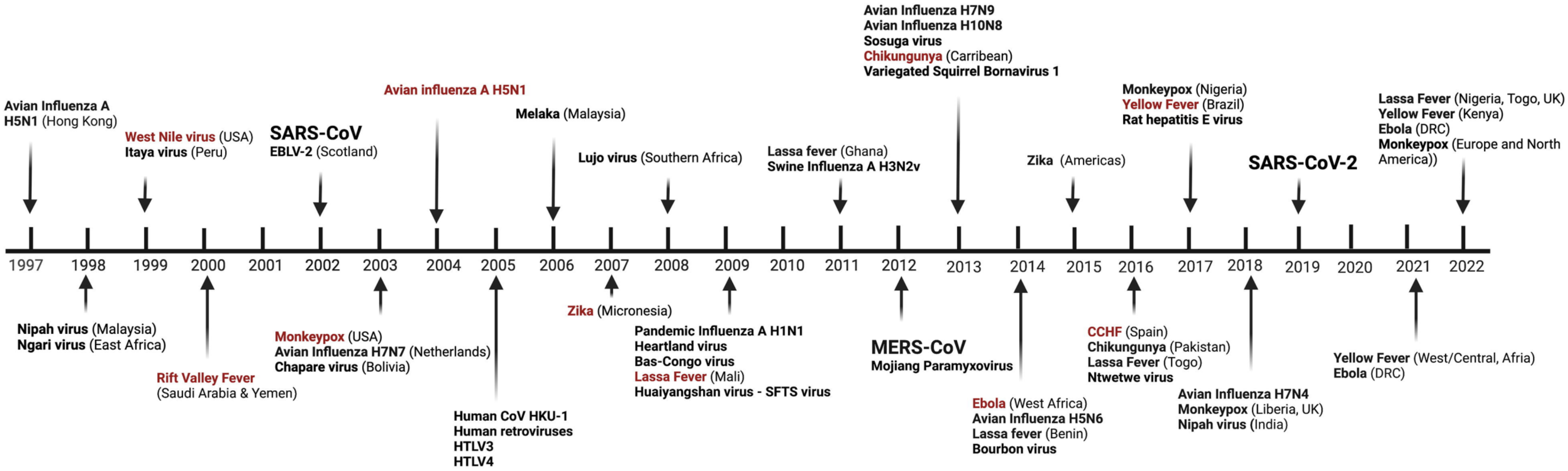
Major zoonotic pandemic outbreaks in the recorded history



Trends in Ecology & Evolution

DOI: https://doi.org/10.1016/j.tree.2020.06.001

Major Emerging and Re-emerging Infectious Disease Outbreaks

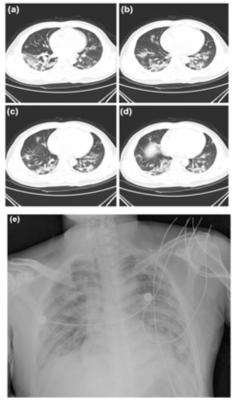


- 1997 – **Avian Influenza A H5N1** (Hong Kong)
- 1998 – **Nipah virus** (Malaysia)
Ngari virus (East Africa)
- 1999 – **West Nile virus** (USA)
Itaya virus (Peru)
- 2000 – **Rift Valley Fever** (Saudi Arabia & Yemen)
- 2001
- 2002 – **SARS-CoV**
EBLV-2 (Scotland)
- 2003 – **Monkeypox** (USA)
Avian Influenza H7N7 (Netherlands)
Chapare virus (Bolivia)
- 2004 – **Avian influenza A H5N1**
- 2005 – **Human CoV HKU-1**
Human retroviruses
HTLV3
HTLV4
- 2006 – **Melaka** (Malaysia)
- 2007 – **Zika** (Micronesia)
- 2008 – **Lujo virus** (Southern Africa)
- 2009 – **Pandemic Influenza A H1N1**
Heartland virus
Bas-Congo virus
Lassa Fever (Mali)
Huaiyangshan virus - SFTS virus
- 2010
- 2011 – **Lassa fever** (Ghana)
Swine Influenza A H3N2v
- 2012 – **MERS-CoV**
Mojiang Paramyxovirus
- 2013 – **Avian Influenza H7N9**
Avian Influenza H10N8
Sosuga virus
Chikungunya (Carribean)
Variegated Squirrel Bornavirus 1
- 2014 – **Ebola** (West Africa)
Avian Influenza H5N6
Lassa fever (Benin)
Bourbon virus
- 2015 – **Zika** (Americas)
- 2016 – **CCHF** (Spain)
Chikungunya (Pakistan)
Lassa Fever (Togo)
Ntwetwe virus
- 2017 – **Monkeypox** (Nigeria)
Yellow Fever (Brazil)
Rat hepatitis E virus
- 2018 – **Avian Influenza H7N4**
Monkeypox (Liberia, UK)
Nipah virus (India)
- 2019 – **SARS-CoV-2**
- 2020
- 2021 – **Yellow Fever** (West/Central, Afria)
Ebola (DRC)
- 2022 – **Lassa Fever** (Nigeria, Togo, UK)
Yellow Fever (Kenya)
Ebola (DRC)
Monkeypox (Europe and North America))

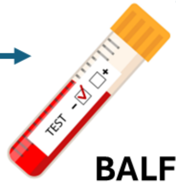
The most problematic EIDs were caused by RNA viruses

Gerald T. Keusch, et al PNAS 2022

Metagenomics in the identification of SARS-CoV-2



a patient who was a worker at the market and was admitted to the Central Hospital of Wuhan on 26 December 2019 while experiencing a severe respiratory syndrome

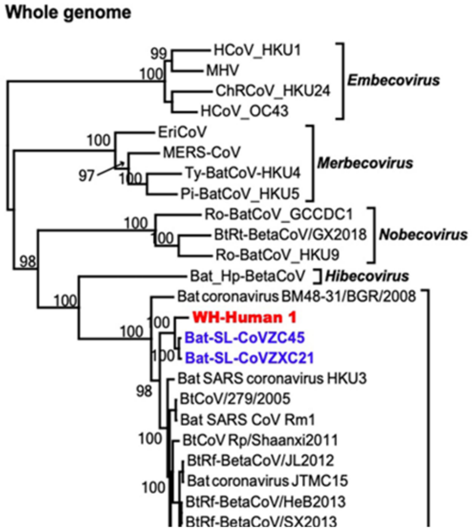


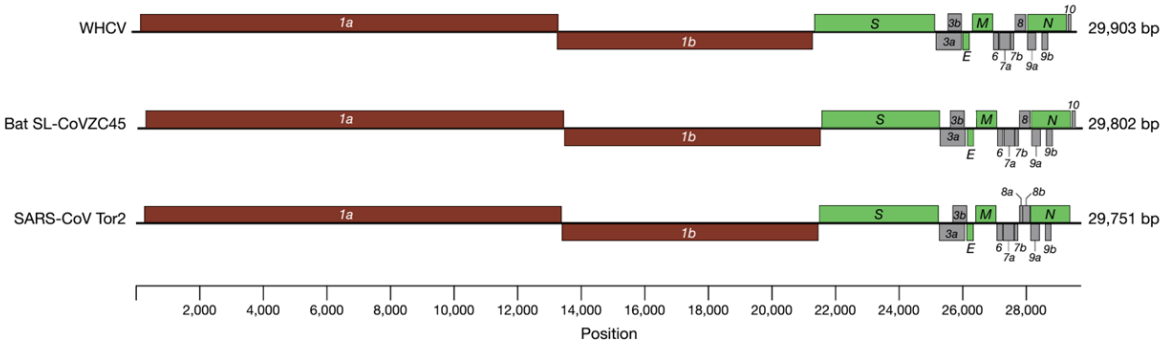
BALF sample

Other respiratory pathogens were not detected by real-time RT–PCR; FluA, FluB, ADV, Chlamydia pneumoniae

Extraction of total RNA, and library preparation

Sequencing with Illumina





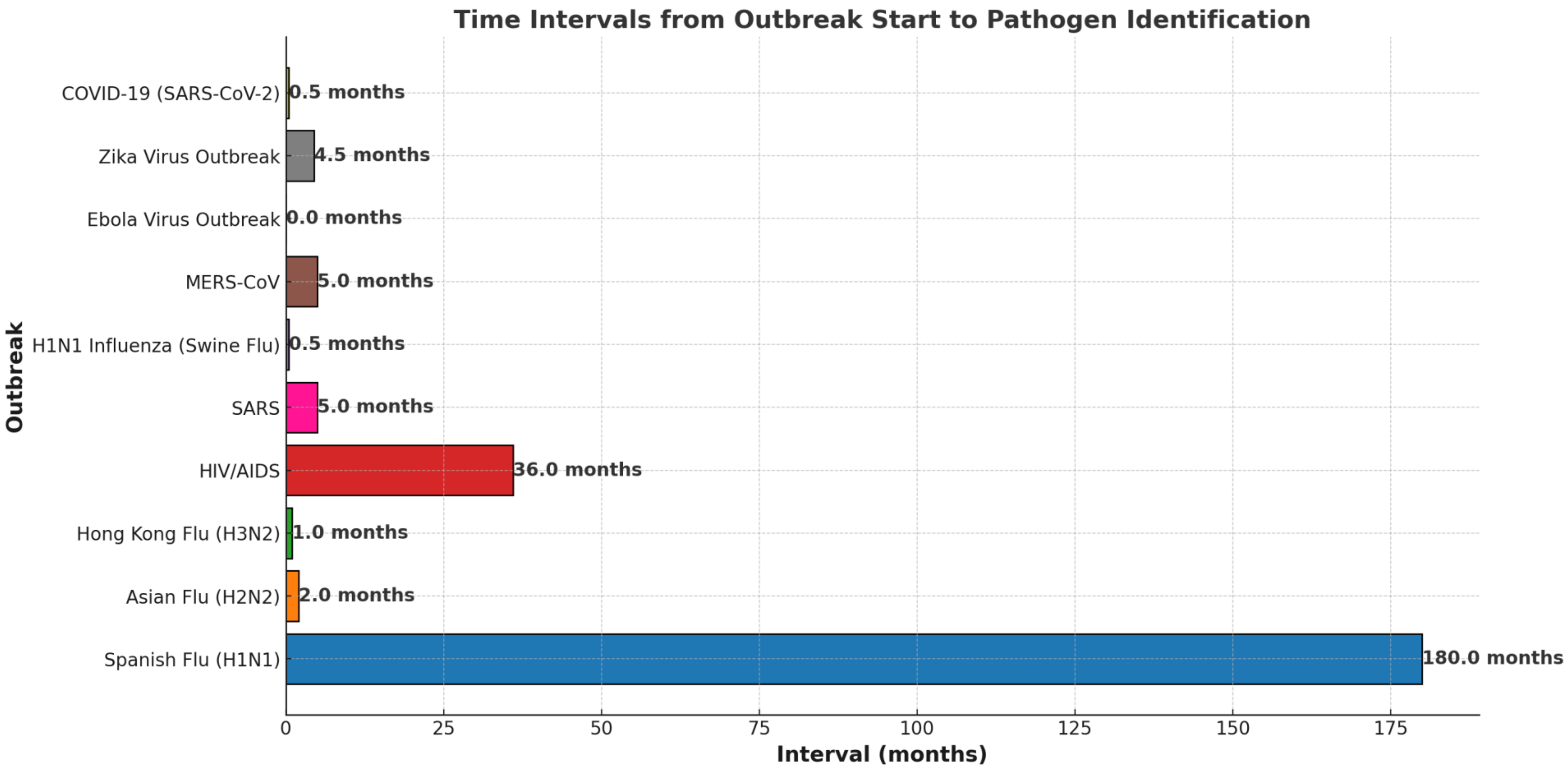
Longest contigs generated by Megahit (30,474 nt) and showed high similarity to the bat SARS-like coronavirus isolate bat SL-CoVZC45

Wu et al, *Nature*, February 2020

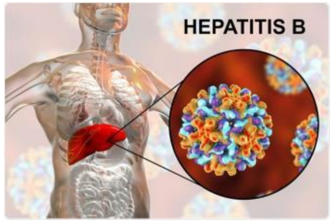
Discovery and identification of SARS-CoV-2

- Metagenomics
- Complete genome sequence
- Primers design for the diagnosis
- Development of mRNA vaccine
- Phylogenetic analysis, Tracking the origin and transmission
- Disease surveillance and control

Interval time between an outbreak and pathogen identification is shorter after mNGS



Milestones in the Discovery of the Hepatitis C Virus

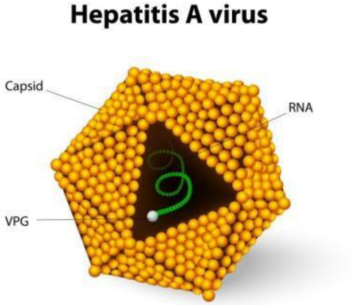


- Australian Antigen discovered

--Two types of hepatitis classified: A and B

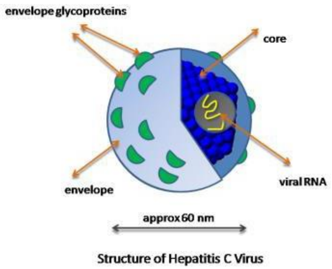
. - Australian antigen linked to hepatitis B

1960



- Hepatitis A identified
- Determination that 75% of TAH caused by NANBH virus
- NANBH transmitted to chimpanzees
- Serologic tests for HAV antibody developed

1970



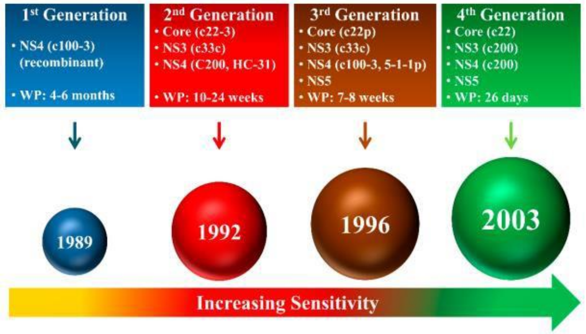
-NANBH discovered to have lipid envelope

-- NANBH discovered to be 30–60 nm

-HCV cloned, sequenced, classified

-Anti-HCV assays developed

1980s



-Use of 1st generation HCV screening assay

-Use of 2nd -generation HCV screening assay

-Cultured the first infectious clone HCV

1990s

Drug development

A curable disease

2000-2020

Harvey James Alter

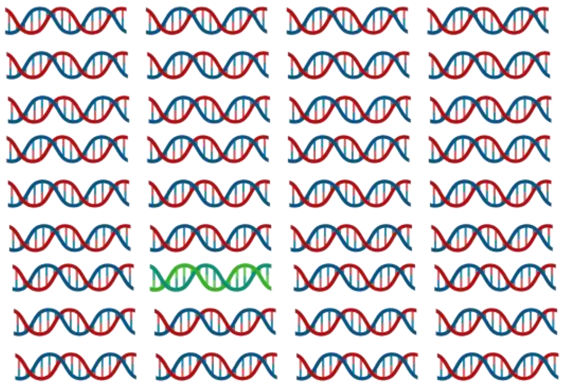
Michael Houghton

Charles Moen Rice

Major challenges when applying NGS in clinical diagnostics

- Complexity in workflows, instruments and data analysis
- Lack of established pipelines for clinical use
- Cost and turnaround time
- Sensitivity and specificity
- Technical challenge
  - Host background
  - Contamination
  - Extraction efficacy

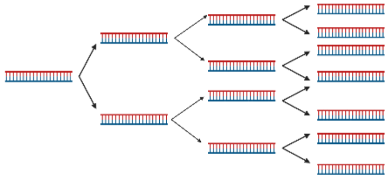
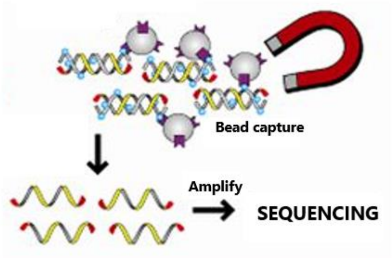
งมเข็ม
ในมหาสมุทร



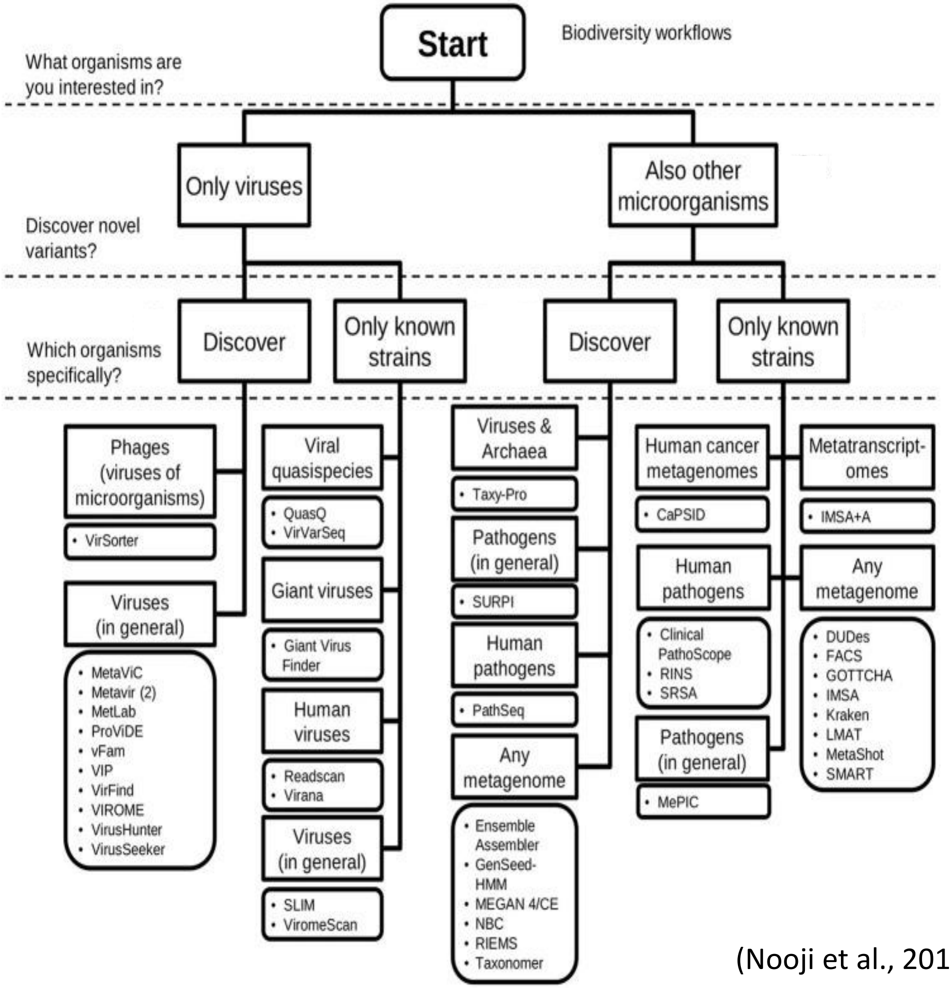
=

>99.9% human related reads

Approaches to reduce host background

1. Nucleic acid depletion	2. Target amplification	3. Host background subtraction	4. Hybridization-based enrichment
			
❖ Physical depletion/ Chemical depletion ❖ Ex: rRNA depletion by probes ❖ Expensive, Deplete homologous non-target sequences, Strand specific bias	❖ Multiplexed PCR ❖ Sequence-independent single primer amplification ❖ Limited number of viruses that can be detected	❖ Computational subtraction of known human sequences ❖ Require high computational power	❖ Use probes covering a multitude of viruses or potentially all viruses ❖ High reproducibility ❖ Comparatively more uniform genome coverage

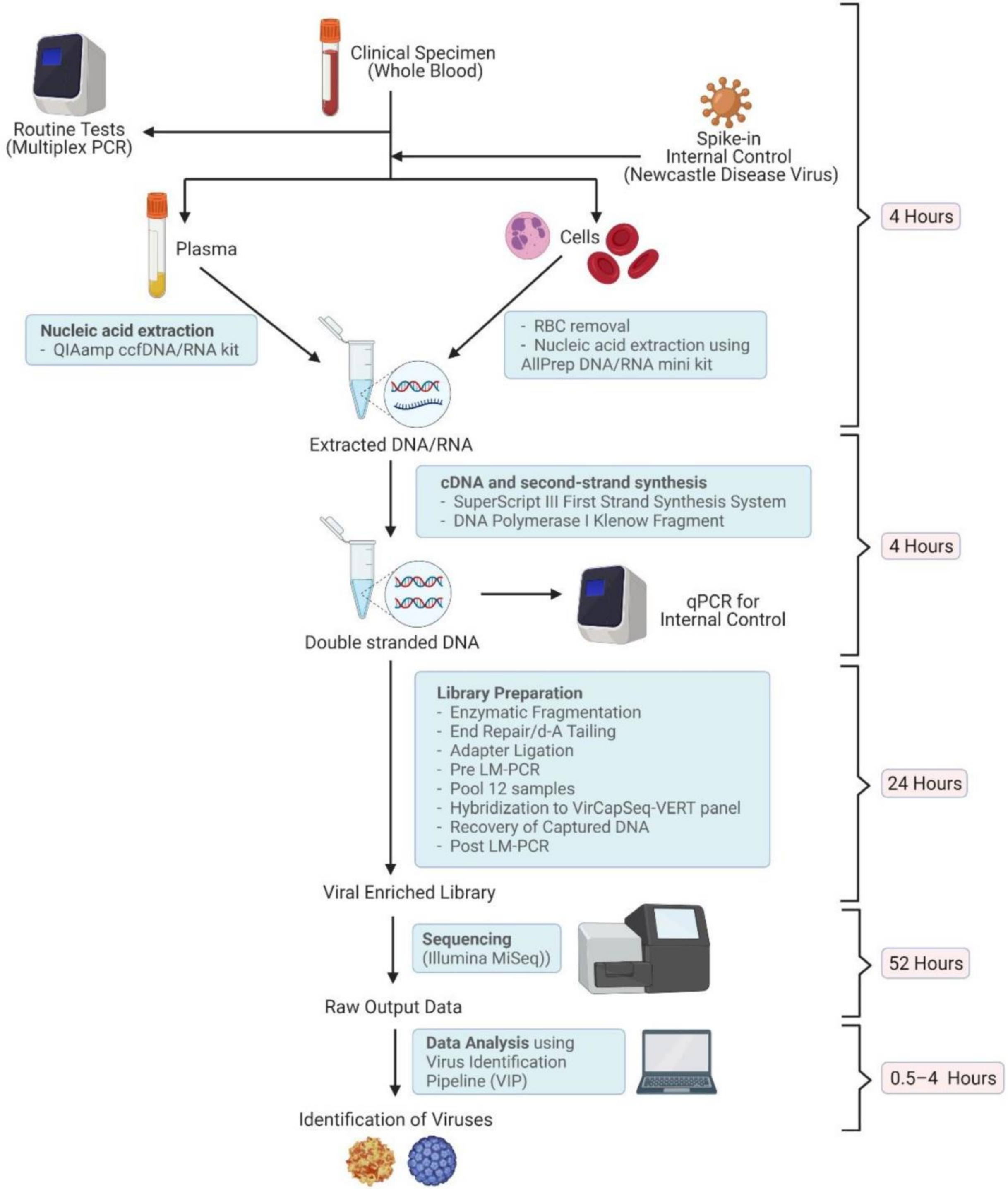
BIOINFORMATICS ANALYSIS



(Nooji et al., 2018)

Tool	Description	Major advantage	Major disadvantage	Link to results
Taxonomer	Characterizes taxonomy of microbiome; web-based	• User-friendly	• Results were not descriptive	https://docs.google.com/spreadsheets/d/1MDgEq_g0TiNwdlLkdc4NGhbLpXWLURNX/edit?usp=sharing&ouid=114885429356739401313&rtpof=true&sd=true
ViromeScan	Characterizes taxonomy of eukaryotic viruses; desktop app	• -	• Only known viruses could be identified / bugs in the software	Could not obtain results
Virus genome-targeted assembly pipeline: **VirusTAP**	Assembly of viral genome sequences; web-based	• User-friendly	• Results were not descriptive	file:///C:/Project%20data/Data%20analysis/VirusTap/VirusTAP_004-D0-2/VirusTAP/web/index.cgi@ID=19033115510ajml&key=0lZnHEHHfLc4KZt5T.html
Virus Identification Pipeline: **VIP**	Identification of eukaryotic viruses from metagenomes; desktop app	• Descriptive & easy to understand results	• Consensus sequence of the identified virus cannot be obtained	file:///C:/Users/nludo/Desktop/Thesis/Raw%20results/VIP_report.html

Human Pegivirus-1 in Pediatric Hematopoietic Stem Cell Transplantation Recipients



Clinical Specimen (Whole Blood)

Routine Tests (Multiplex PCR)

Spike-in Internal Control (Newcastle Disease Virus)

Plasma

Cells

Nucleic acid extraction
- QIAamp ccfDNA/RNA kit

- RBC removal
- Nucleic acid extraction using AllPrep DNA/RNA mini kit

Extracted DNA/RNA

4 Hours

cDNA and second-strand synthesis
- SuperScript III First Strand Synthesis System
- DNA Polymerase I Klenow Fragment

Double stranded DNA

qPCR for Internal Control

4 Hours

Library Preparation
- Enzymatic Fragmentation
- End Repair/d-A Tailing
- Adapter Ligation
- Pre LM-PCR
- Pool 12 samples
- Hybridization to VirCapSeq-VERT panel
- Recovery of Captured DNA
- Post LM-PCR

Viral Enriched Library

24 Hours

Sequencing (Illumina MiSeq))

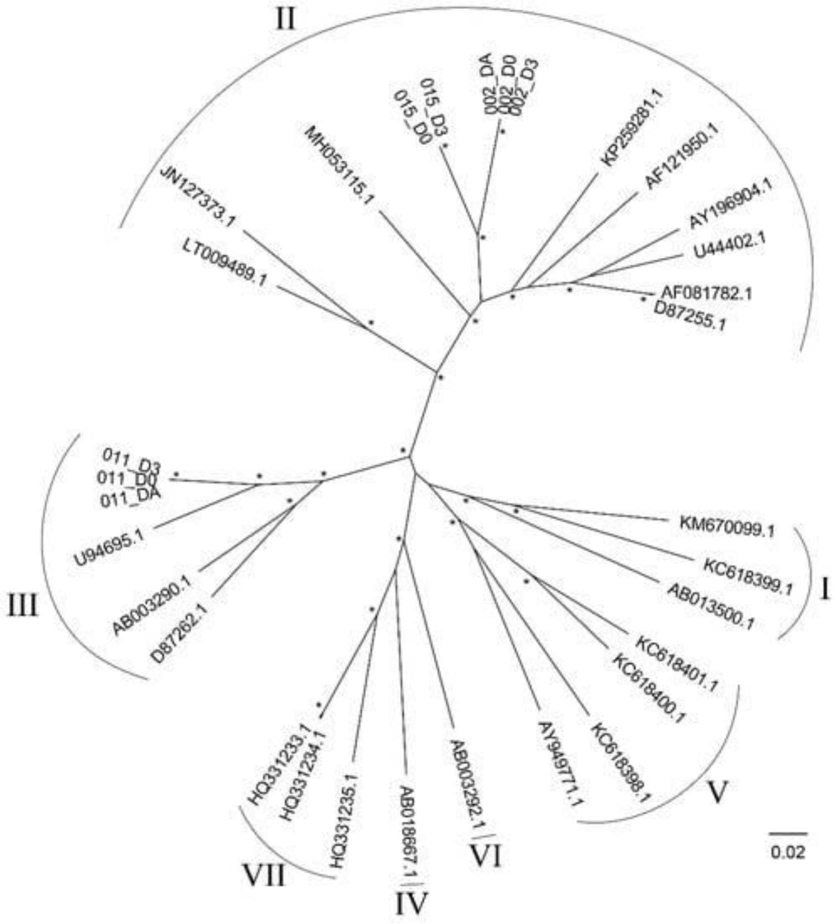
Raw Output Data

52 Hours

Data Analysis using Virus Identification Pipeline (VIP)

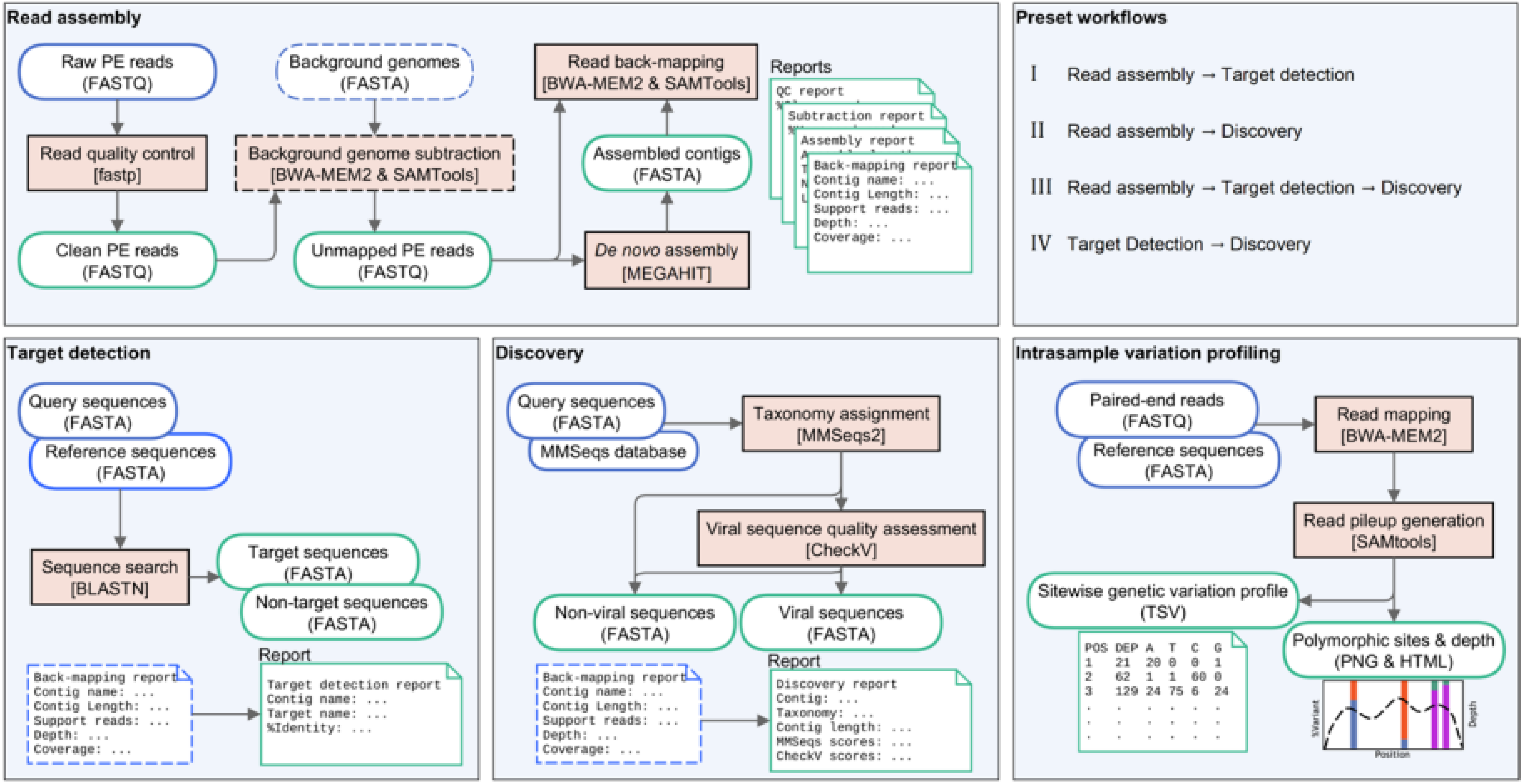
Identification of Viruses

0.5–4 Hours



Natali Ludowyke,et al Viruses 2022, 14(4), 796

Entourage: all-in-one sequence analysis, software for genome assembly, virus detection, virus discovery, and intrasample variation profiling



Read assembly

Raw PE reads (FASTQ)

Read quality control [fastp]

Clean PE reads (FASTQ)

Background genomes (FASTA)

Background genome subtraction [BWA-MEM2 & SAMTools]

Unmapped PE reads (FASTQ)

De novo assembly [MEGAHIT]

Assembled contigs (FASTA)

Read back-mapping [BWA-MEM2 & SAMTools]

Reports

```
QC report
Subtraction report
Assembly report
Back-mapping report
Contig name: ...
Contig Length: ...
Support reads: ...
Depth: ...
Coverage: ...
```

Preset workflows

I Read assembly → Target detection

II Read assembly → Discovery

III Read assembly → Target detection → Discovery

IV Target Detection → Discovery

Target detection

Query sequences (FASTA)

Reference sequences (FASTA)

Sequence search [BLASTN]

Target sequences (FASTA)

Non-target sequences (FASTA)

```
Back-mapping report
Contig name: ...
Contig Length: ...
Support reads: ...
Depth: ...
Coverage: ...
```

Report

```
Target detection report
Contig name: ...
Target name: ...
%Identity: ...
```

Discovery

Query sequences (FASTA)

MMSeqs database

Taxonomy assignment [MMSeqs2]

Viral sequence quality assessment [CheckV]

Non-viral sequences (FASTA)

Viral sequences (FASTA)

```
Back-mapping report
Contig name: ...
Contig Length: ...
Support reads: ...
Depth: ...
Coverage: ...
```

Report

```
Discovery report
Contig: ...
Taxonomy: ...
Contig length: ...
MMSeqs scores: ...
CheckV scores: ...
```

Intrasample variation profiling

Paired-end reads (FASTQ)

Reference sequences (FASTA)

Read mapping [BWA-MEM2]

Read pileup generation [SAMtools]

Sitewise genetic variation profile (TSV)

```
POS  DEP  A   T   C   G
1    21   20  0   0   1
2    62   1   1   60  0
3    129  24  75  6   24
.    .    .   .   .   .
.    .    .   .   .   .
.    .    .   .   .   .
```

Polymorphic sites & depth (PNG & HTML)

https://codeberg.org/CENMIG/Entourage

Phumiphanjarphak and Aiewsakun BMC Bioinformatics (2024) 25:222

Entourage’s features compared to other currently available virus discovery software

Feature	Entourage	VirusSeeker [13]	SURPI [14]	GATK PathSeq [15]	VirFind [16]	VIP [17]	Lazypipe [18]	Genome Detective [19]	VIRify [20]
Functionalities									
Read quality control	✓	✓	✓	✓	✓	✓	✓	✓	
Background sequence subtraction	✓	✓	✓	✓	✓	✓	✓		
Read binning		✓	✓	✓		✓		✓	
Sequence assembly:	✓	✓	✓		✓	✓	✓	✓	
pre-taxonomic identification	✓	✓	✓		✓		✓		
post-taxonomic identification						✓		✓	
Target detection	✓			✓					✓
Virus discovery	✓	✓	✓	✓	✓	✓	✓	✓	✓
Intrasample variation profiling	✓								
Input flexibility									
Use raw reads as input	✓	✓	✓	✓	✓	✓	✓	✓	
Use assembled sequence as input	✓							✓	✓
Software execution									
Batch processing	✓	✓							
Local execution	✓	✓	✓	✓		✓	✓		✓

Phumiphanjarphak and Aiewsakun BMC Bioinformatics (2024) 25:222

Several pipelines have been developed to support viral detection and discovery

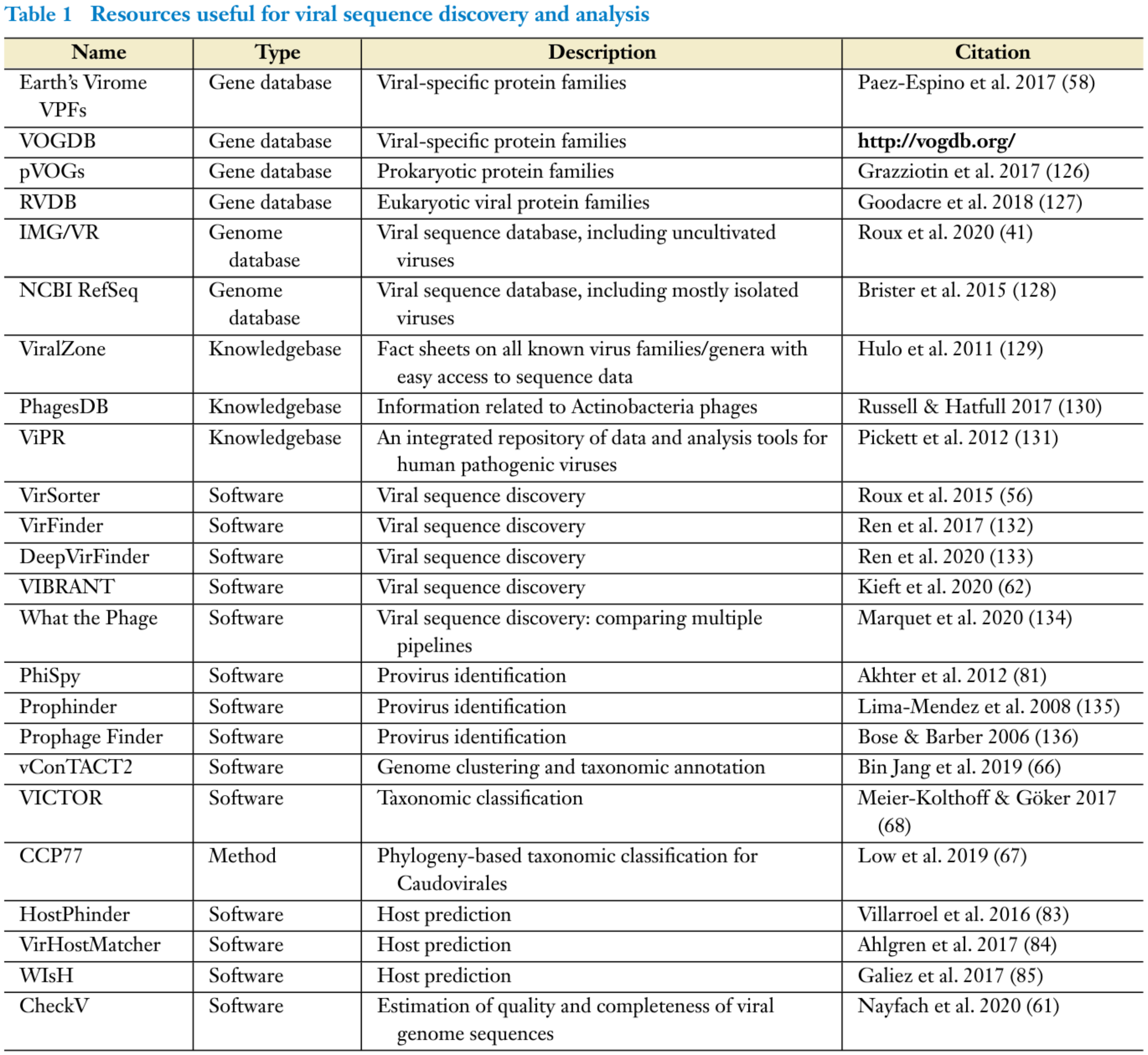
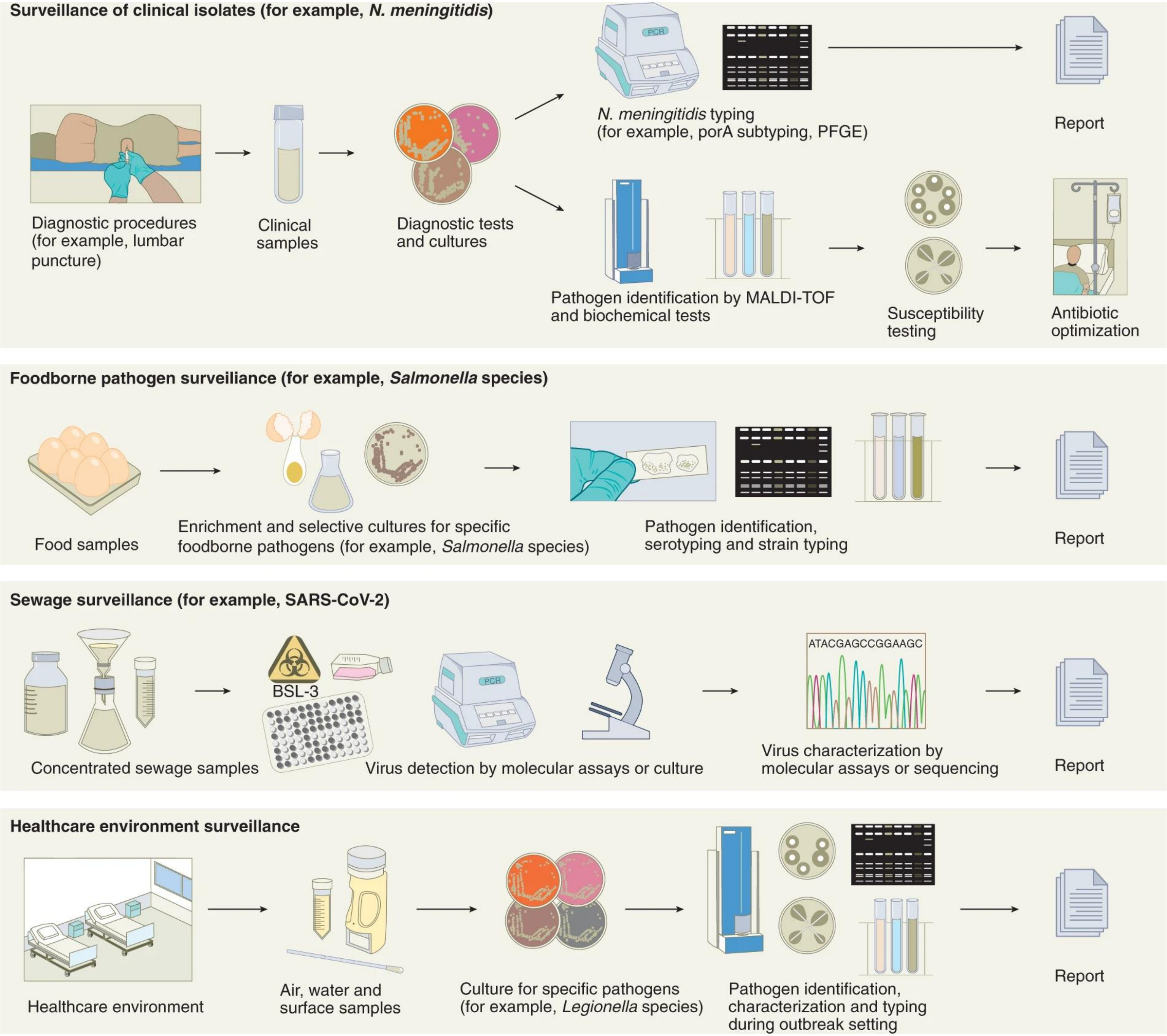


Table 1 Resources useful for viral sequence discovery and analysis

Name	Type	Description	Citation
Earth's Virome VPFs	Gene database	Viral-specific protein families	Paez-Espino et al. 2017 (58)
VOGDB	Gene database	Viral-specific protein families	**http://vogdb.org/**
pVOGs	Gene database	Prokaryotic protein families	Grazziotin et al. 2017 (126)
RVDB	Gene database	Eukaryotic viral protein families	Goodacre et al. 2018 (127)
IMG/VR	Genome database	Viral sequence database, including uncultivated viruses	Roux et al. 2020 (41)
NCBI RefSeq	Genome database	Viral sequence database, including mostly isolated viruses	Brister et al. 2015 (128)
ViralZone	Knowledgebase	Fact sheets on all known virus families/genera with easy access to sequence data	Hulo et al. 2011 (129)
PhagesDB	Knowledgebase	Information related to Actinobacteria phages	Russell & Hatfull 2017 (130)
ViPR	Knowledgebase	An integrated repository of data and analysis tools for human pathogenic viruses	Pickett et al. 2012 (131)
VirSorter	Software	Viral sequence discovery	Roux et al. 2015 (56)
VirFinder	Software	Viral sequence discovery	Ren et al. 2017 (132)
DeepVirFinder	Software	Viral sequence discovery	Ren et al. 2020 (133)
VIBRANT	Software	Viral sequence discovery	Kieft et al. 2020 (62)
What the Phage	Software	Viral sequence discovery: comparing multiple pipelines	Marquet et al. 2020 (134)
PhiSpy	Software	Provirus identification	Akhter et al. 2012 (81)
Prophinder	Software	Provirus identification	Lima-Mendez et al. 2008 (135)
Prophage Finder	Software	Provirus identification	Bose & Barber 2006 (136)
vConTACT2	Software	Genome clustering and taxonomic annotation	Bin Jang et al. 2019 (66)
VICTOR	Software	Taxonomic classification	Meier-Kolthoff & Göker 2017 (68)
CCP77	Method	Phylogeny-based taxonomic classification for Caudovirales	Low et al. 2019 (67)
HostPhinder	Software	Host prediction	Villarroel et al. 2016 (83)
VirHostMatcher	Software	Host prediction	Ahlgren et al. 2017 (84)
WIsH	Software	Host prediction	Galiez et al. 2017 (85)
CheckV	Software	Estimation of quality and completeness of viral genome sequences	Nayfach et al. 2020 (61)

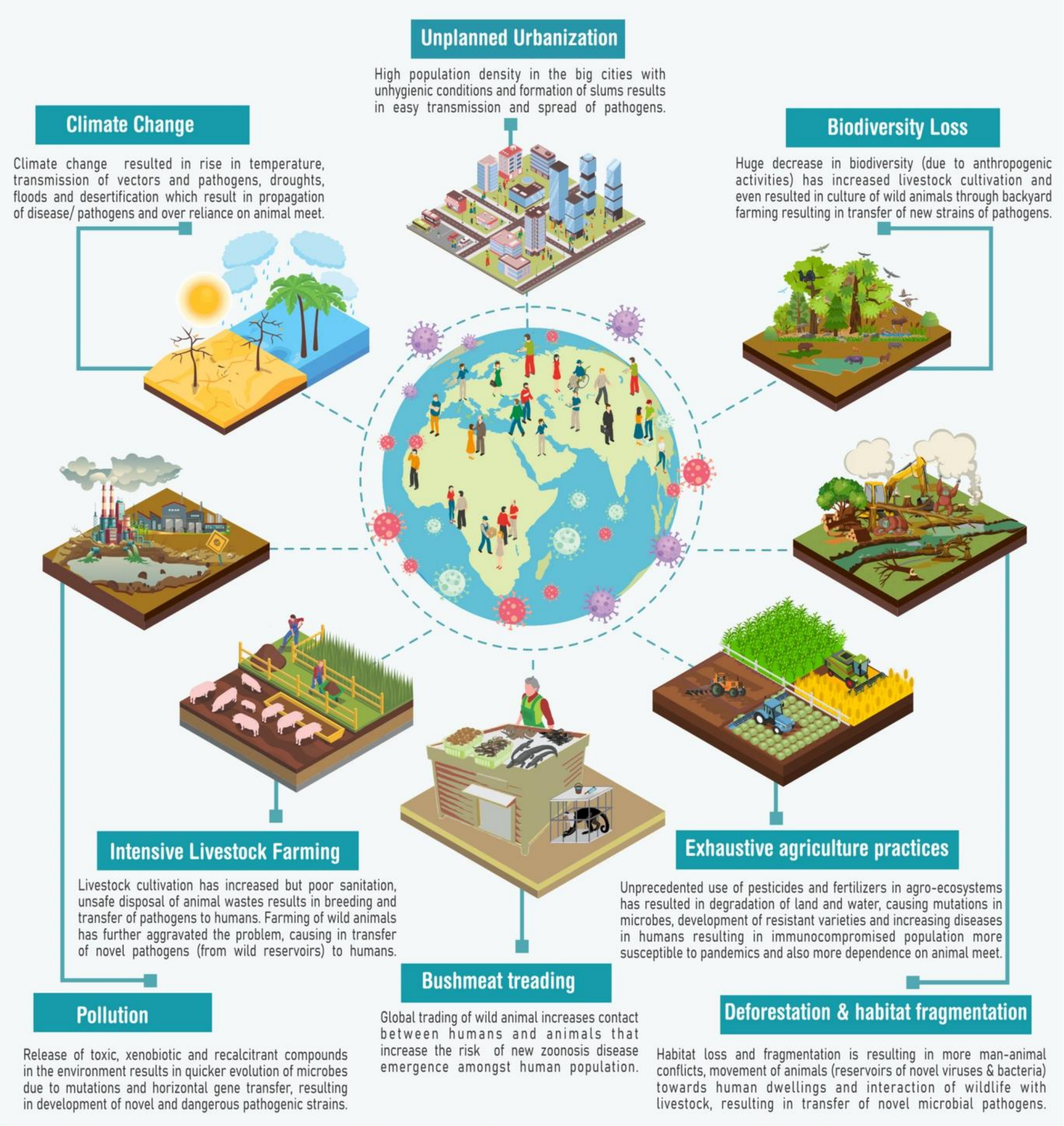
Call, Nayfach, and Kyrpides, *Annual Review of Biomedical Data Science*, 2021

Metagenomics-enabled microbial surveillance



Nature Microbiology volume 7, pages486–496 (2022)

Environmental issues that lead to the pandemic



Unplanned Urbanization

High population density in the big cities with unhygienic conditions and formation of slums results in easy transmission and spread of pathogens.

Climate Change

Climate change resulted in rise in temperature, transmission of vectors and pathogens, droughts, floods and desertification which result in propagation of disease/ pathogens and over reliance on animal meet.

Biodiversity Loss

Huge decrease in biodiversity (due to anthropogenic activities) has increased livestock cultivation and even resulted in culture of wild animals through backyard farming resulting in transfer of new strains of pathogens.

Intensive Livestock Farming

Livestock cultivation has increased but poor sanitation, unsafe disposal of animal wastes results in breeding and transfer of pathogens to humans. Farming of wild animals has further aggravated the problem, causing in transfer of novel pathogens (from wild reservoirs) to humans.

Exhaustive agriculture practices

Unprecedented use of pesticides and fertilizers in agro-ecosystems has resulted in degradation of land and water, causing mutations in microbes, development of resistant varieties and increasing diseases in humans resulting in immunocompromised population more susceptible to pandemics and also more dependence on animal meet.

Pollution

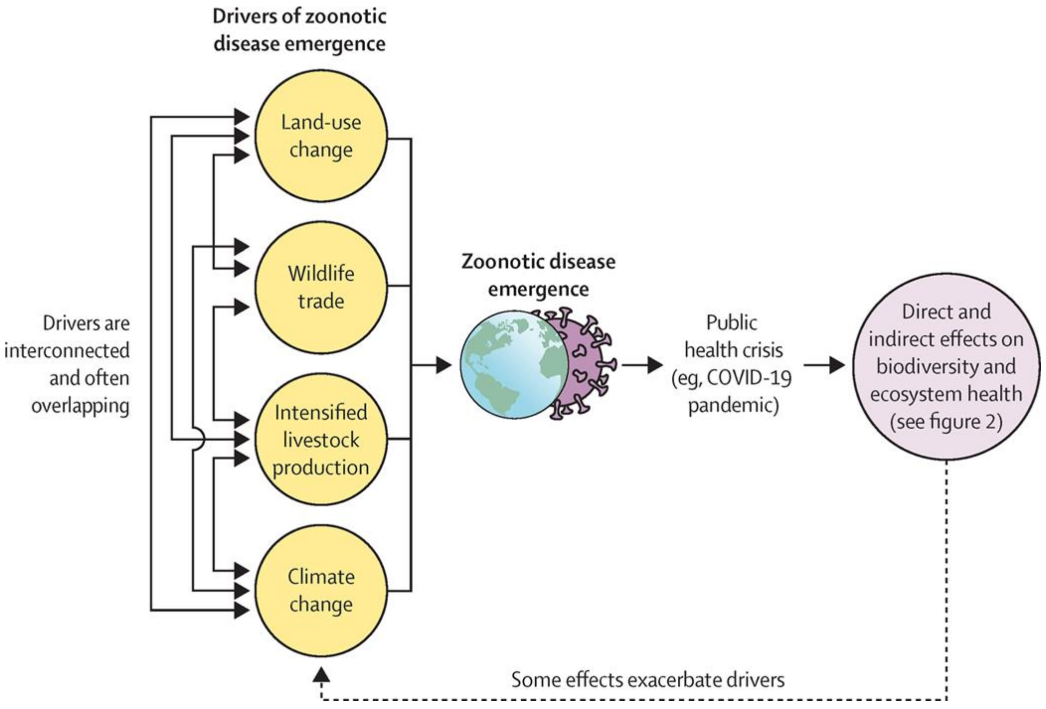
Release of toxic, xenobiotic and recalcitrant compounds in the environment results in quicker evolution of microbes due to mutations and horizontal gene transfer, resulting in development of novel and dangerous pathogenic strains.

Bushmeat treading

Global trading of wild animal increases contact between humans and animals that increase the risk of new zoonosis disease emergence amongst human population.

Deforestation & habitat fragmentation

Habitat loss and fragmentation is resulting in more man-animal conflicts, movement of animals (reservoirs of novel viruses & bacteria) towards human dwellings and interaction of wildlife with livestock, resulting in transfer of novel microbial pathogens.



Drivers of zoonotic disease emergence

Land-use change

Wildlife trade

Intensified livestock production

Climate change

Drivers are interconnected and often overlapping

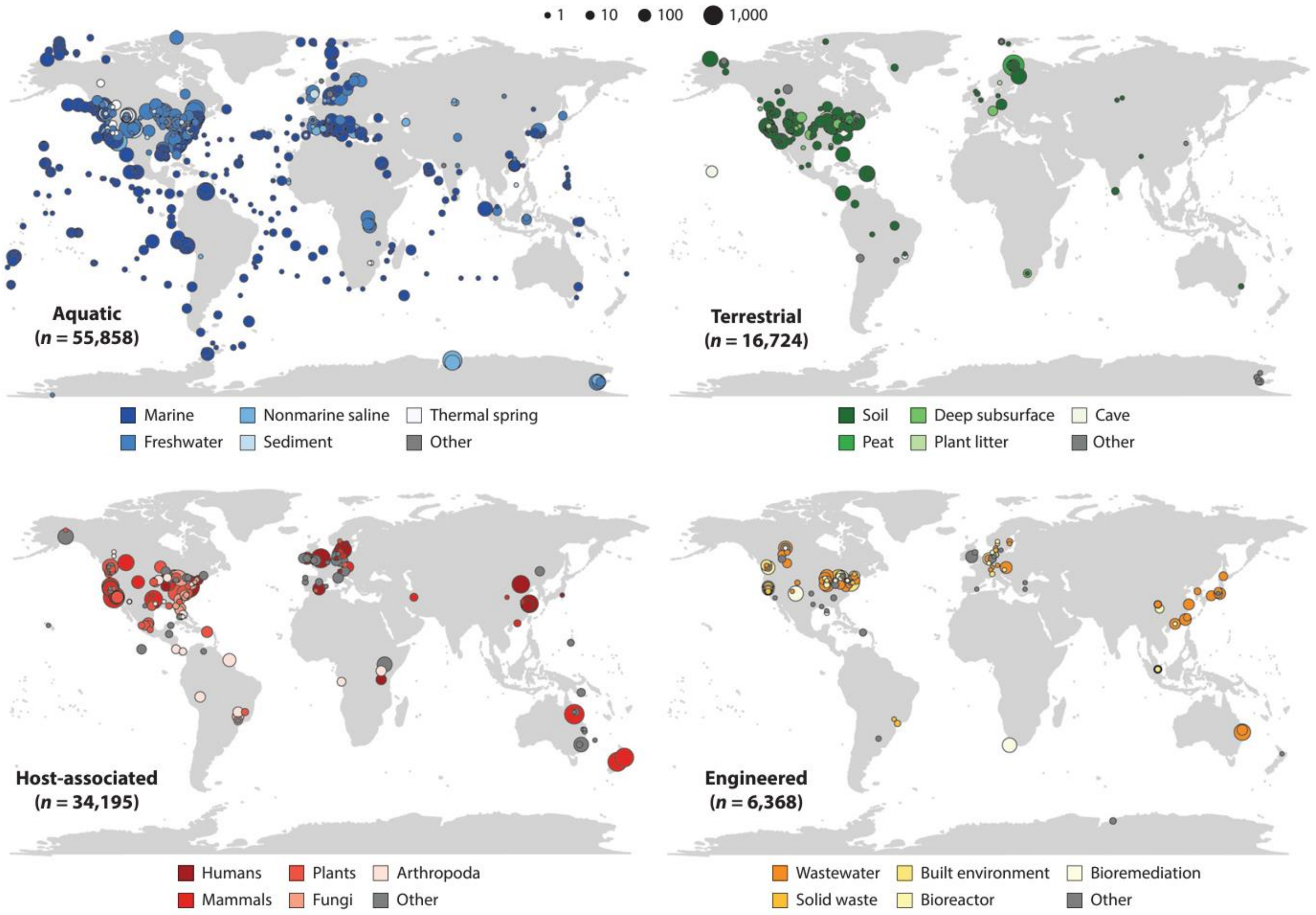
Zoonotic disease emergence

Public health crisis (eg, COVID-19 pandemic)

Direct and indirect effects on biodiversity and ecosystem health (see figure 2)

Some effects exacerbate drivers

Distribution of metagenome-assembled viral genomes across Earth’s biomes



Call, Nayfach, and Kyrpides, *Annual Review of Biomedical Data Science*, 2021

Potential Solutions and Future Directions

- Improved Methods
- Comprehensive Databases
- Interdisciplinary Collaboration
- Standardization

Bioinformatics tools should be automated and freely available to allow labs to perform viral metagenomic analyses.

Summary

- Clinical metagenomics is slowly becoming part of today's clinician’s tool to identify infectious diseases.
- Challenges in NGS are being overcome in clinical microbiology laboratories.
- Many challenges remain to the routine use and implementation of these methods.
- Importance of interdisciplinary collaboration for success.

‘Next pandemic could be more lethal than Covid’

Sarah Gilbert

Acknowledgment

- Center for Medical Genomics
- Virus Unit RAMA
- Pediatric Infectious diseases
- Pediatric Hematology/Oncology
- Faculty of Science
- CENMIG