bioinformatics.ca

Canadian Bioinformatics Workshops

www.bioinformatics.ca

bioinformaticsdotca.github.io

CC BY-SA 4.0 DEED

Attribution-ShareAlike 4.0 International

Module 1

Introduction to Genomic Epidemiology

William Hsiao

Infectious Disease Genomic Epidemiology

May 13-17, 2024

bioinformatics.ca

CENTRE FOR INFECTIOUS DISEASE GENOMICS AND ONE HEALTH

SFU SIMON FRASER UNIVERSITY

Course Overview

- Module 1: **Introduction to Genomic Epidemiology**
- Module 2: **De Novo Genome Assembly and Annotation (Gary Van DomSelaar)**
- Module 3: **Data Curation and Data Sharing (Emma Griffiths)**
- Module 4: **Phylogenetic Analysis (Fiona Brinkman)**
- Module 5: **Viral Pathogen Genomic Analysis (Jared Simpson)**
- Module 6: **Bacterial Pathogen Genomic Analysis (Ed Taboada)**
- Module 7: **Antimicrobial Resistant Gene Analysis (Andrew McArthur)**
- Module 8: **Phylodynamics (Finlay Maguire)**
- Module 9: **Mobile Genetic Elements (Rob Beiko)**
- Module 10: **Emerging Pathogen Detection and Identification (Darian Hole)**
- **+ Integrated Assignment**

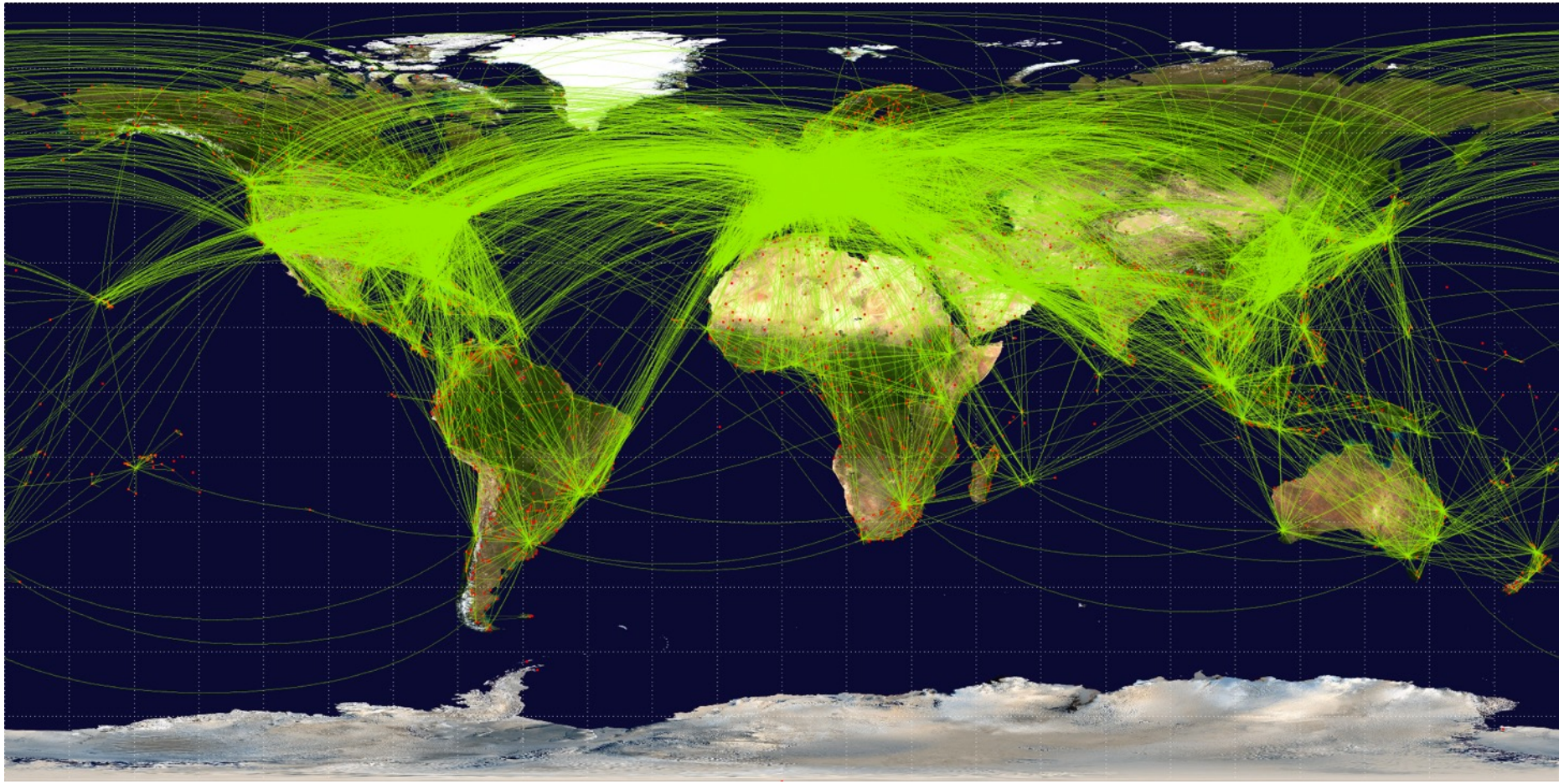
General Learning Objectives

- By the end of this workshop, you will:
  - Understand how genomic epidemiology can improve clinical and public health microbiology
  - Process genomic sequence data using a variety of bioinformatics tools for bacterial and viral genomes and metagenomes
  - Interpret genomic data in different epidemiological contexts and understand the importance of data standardization and sharing
  - Perform several types of genomic epidemiology analyses
  - Recognize the limitations and challenges of genomic epidemiology (a rapidly evolving field)

Learning Objectives of Module 1

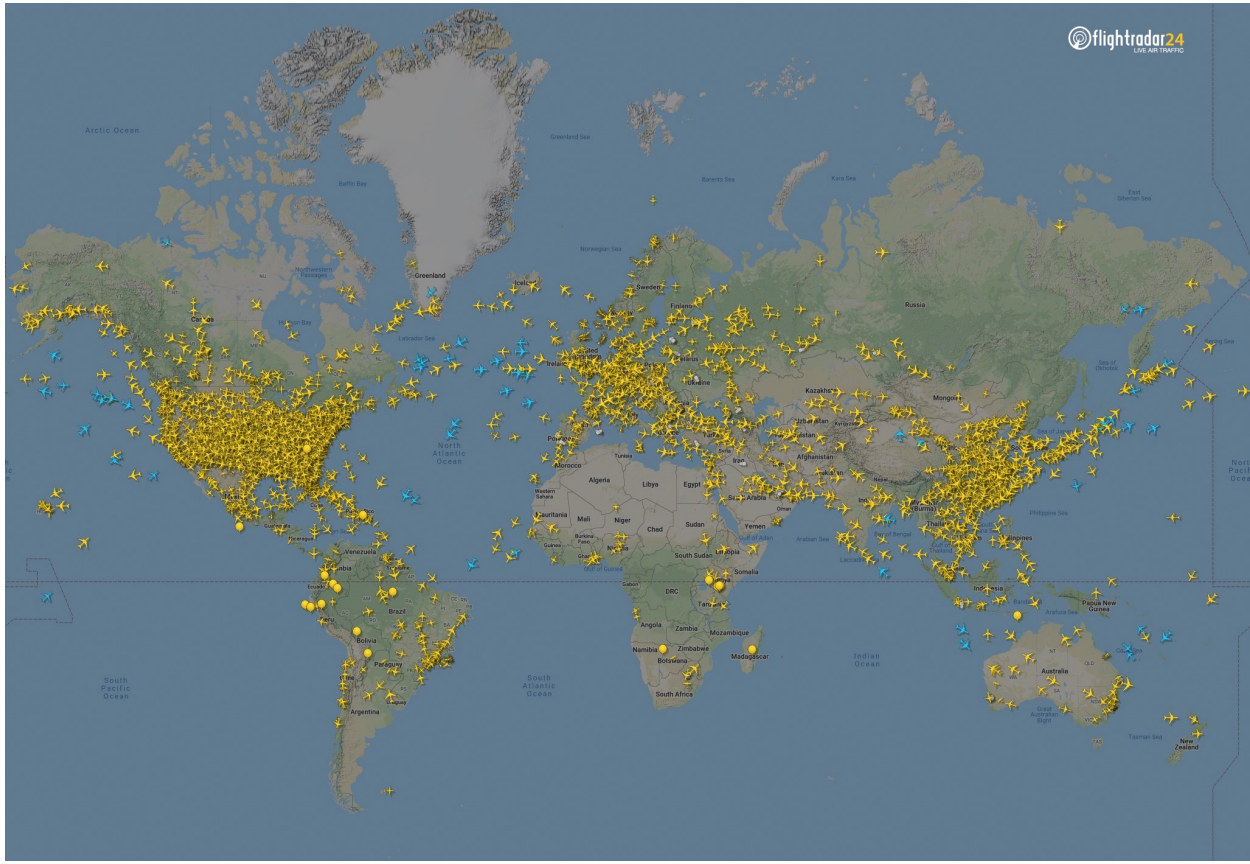
- Understand why infectious disease research is important
- Be familiar with some examples of genomic epidemiology studies
- Be familiar with high-throughput sequencing and its application to clinical and public health microbiology
- Be familiar with sequence data processing steps (more in module 2)
- Understand the challenges associated with sharing genomic epidemiology data (more in module 3)

Global Flight Paths



http://openflights.org/data.html

March 7 2020 v.s. April 7 2020



[Then and now: visualizing COVID-19's impact on air traffic | Flightradar24 Blog]

OPEN

Meta-genomic analysis of toilet waste from long distance flights; a step towards global surveillance of infectious diseases and antimicrobial resistance

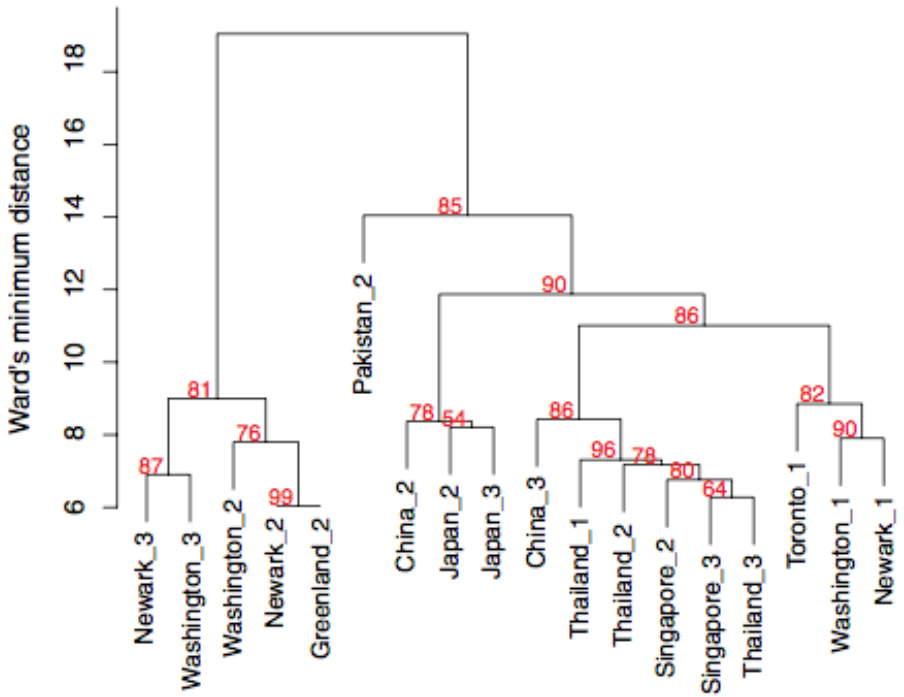
Received: 17 December 2014
Accepted: 17 April 2015
Published: 10 July 2015

Thomas Nordahl Petersen[1], Simon Rasmussen[1], Henrik Hasman[2], Christian Carøe[1], Jacob Bælum[1], Anna Charlotte Schultz[2], Lasse Bergmark[2], Christina A. Svendsen[2], Ole Lund[1], Thomas Sicheritz-Pontén[1] & Frank M. Aarestrup[2]

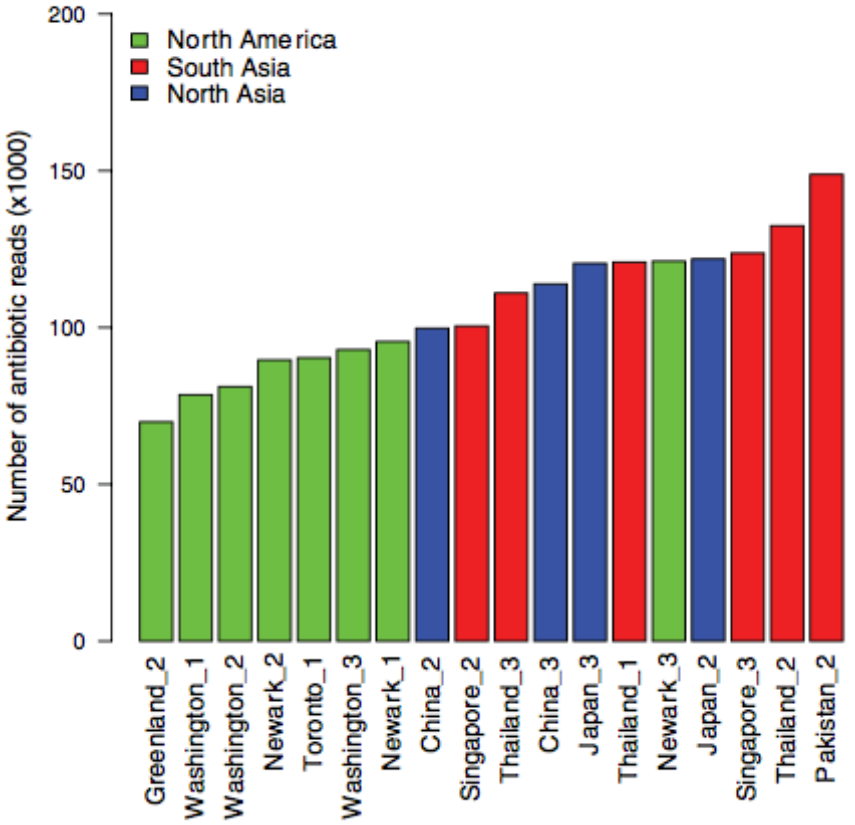
- 18 flights from 3 continents
- Onboard human wastes (400L per flight!) extracted for DNA sequencing

- Samples clustered based on microbiome profile
- Antimicrobial resistance genes identified (~0.06% of the reads)

Petersen, T. N. *et al.* Meta-genomic analysis of toilet waste from long distance flights; a step towards global surveillance of infectious diseases and antimicrobial resistance. *Sci Rep* 1–9 (2015). doi:10.1038/srep11444



Clustering by geographic origins



Higher proportions of antibiotic resistance genes found in flights from South Asia

Petersen, T. N. *et al.* Meta-genomic analysis of toilet waste from long distance flights; a step towards global surveillance of infectious diseases and antimicrobial resistance. *Sci Rep* 1–9 (2015). doi:10.1038/srep11444

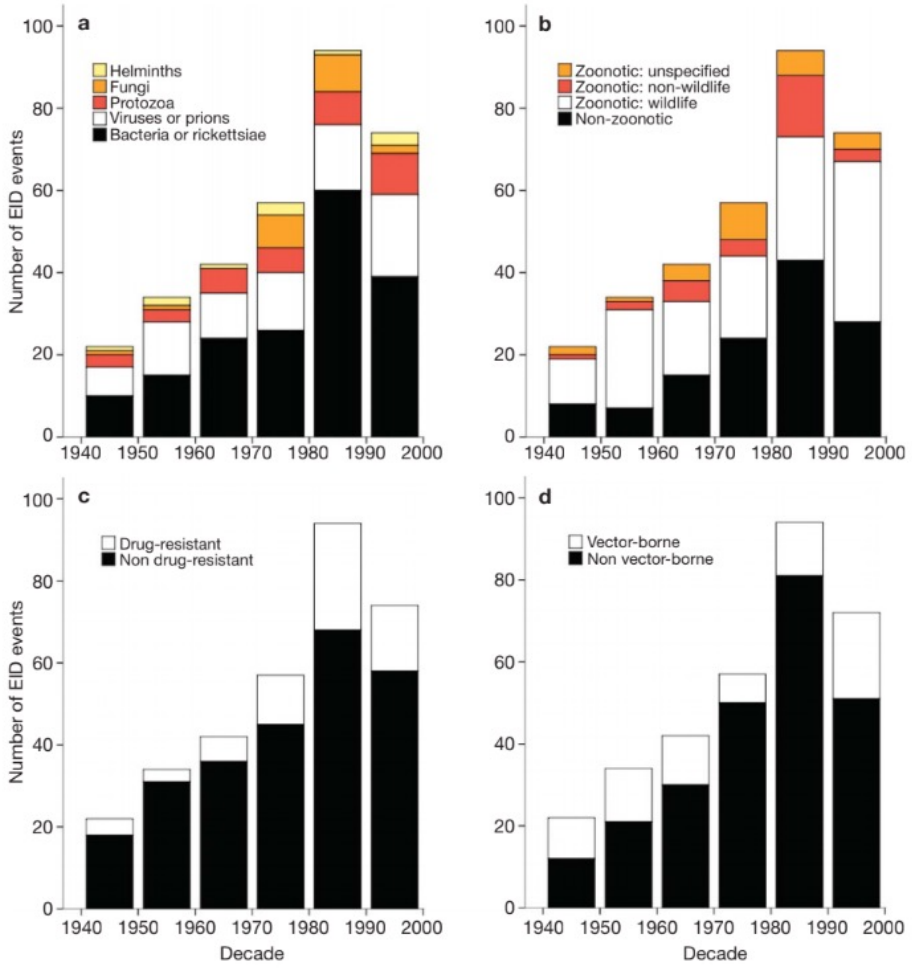
Increase in Emerging Infectious Diseases

EID events: detection of newly evolved strains of pathogens in human

Dominated by zoonotic diseases

Identified global hotspots for EID events

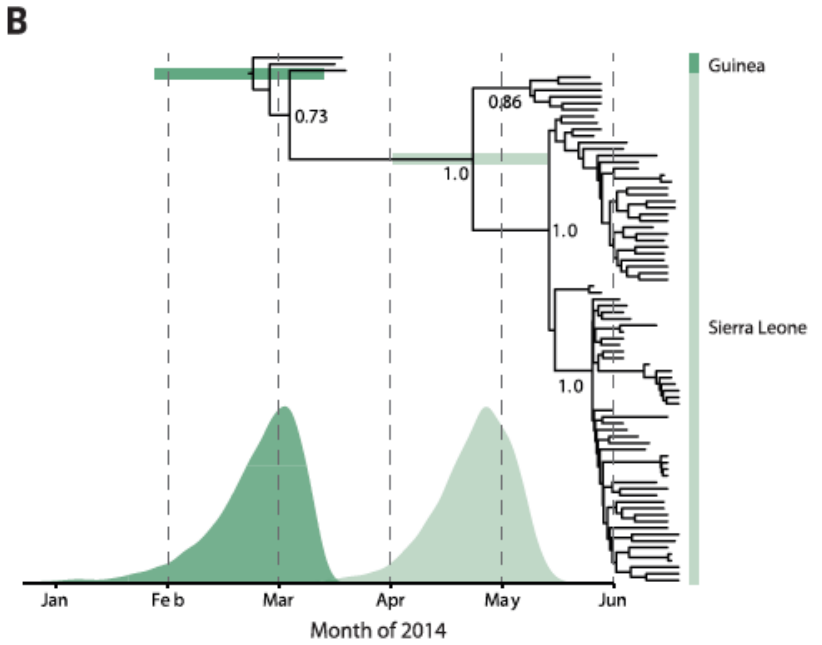
Identified risk factors for EIDs: tropical rain forest, population density, climate, mammal species richness, etc.



Jones. et al 2008. “Global Trends in Emerging Infectious Diseases.” *Nature* 451 (7181): 990–93.
Allen. 2017. “Global Hotspots and Correlates of Emerging Zoonotic Diseases.” *Nature Communications* 8 (1): 1124.

Genomic Sequence Analysis of Ebola

- Outbreak in West Africa from Dec 2013 - May 2016 resulted in 28,616 reported cases and 11,310 deaths
- Global impact on travel
- Genomic epidemiological analysis of 99 early isolates revealed:
  - A single human exposure to natural reservoir (likely bats)
  - Outbreak sustained by **human to human transmission** (e.g. funeral practice and lack of proper quarantine facilities
  - Transmission from Guinea to Sierra Leone likely to be from a single event but two distinct lineages



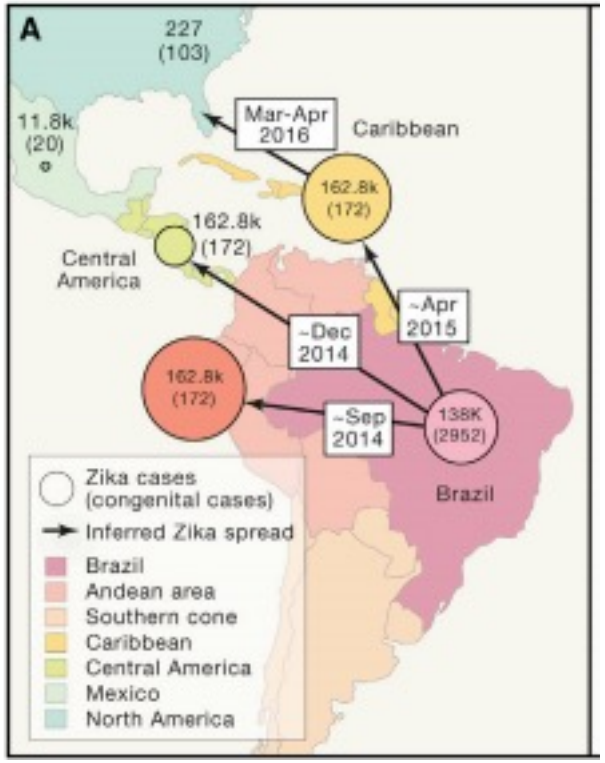
- Sequence data corroborated with epidemiological narrative critical to unravel the complex situation and institute effective policies and interventions... Still there were significant delays

http://www.who.int/csr/disease/ebola/en/

Gire et al. 2014. *Science* 345 (6202): 1369–72.

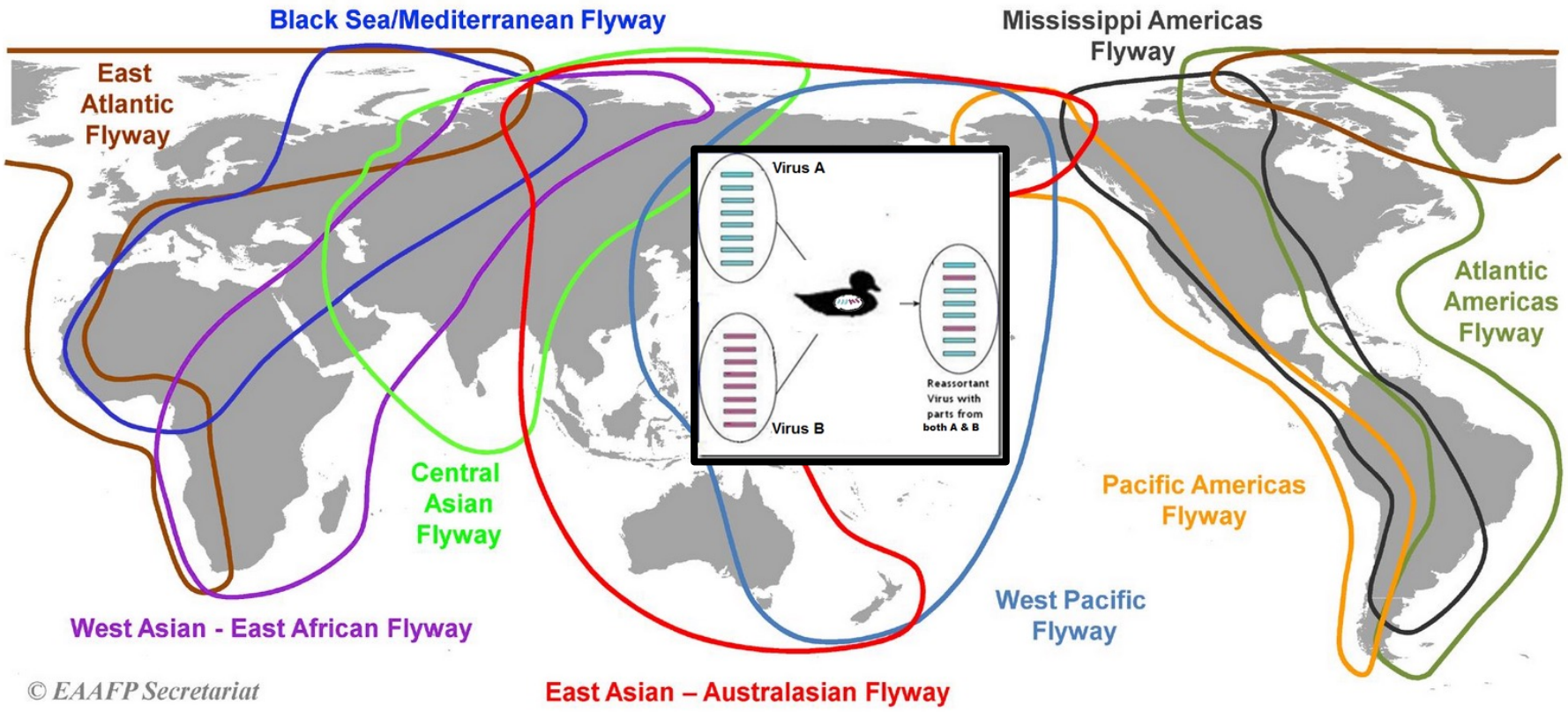
Genomic Insights into Zika Outbreak

- Endemic in Africa and Asia; mild symptoms; self-eliminating
- Caused an outbreak in Americas in 2015-2016; naïve population; high number of people infected and resulted in a few thousand microcephaly cases
- Due to symptoms overlapping with other viruses (e.g. dengue virus), syndromic surveillance can be un-reliable
  - Serological or genomic/molecular tests needed for confirmation
- Phylogenetic reconstruction using genomic data highlighted unsuspected circulation of zika in the population prior to outbreak and help to reconstruct the transmission route



Grubaugh. 2018. “Genomic Insights into Zika Virus Emergence and Spread.” *Cell* 172 (6): 1160–62.

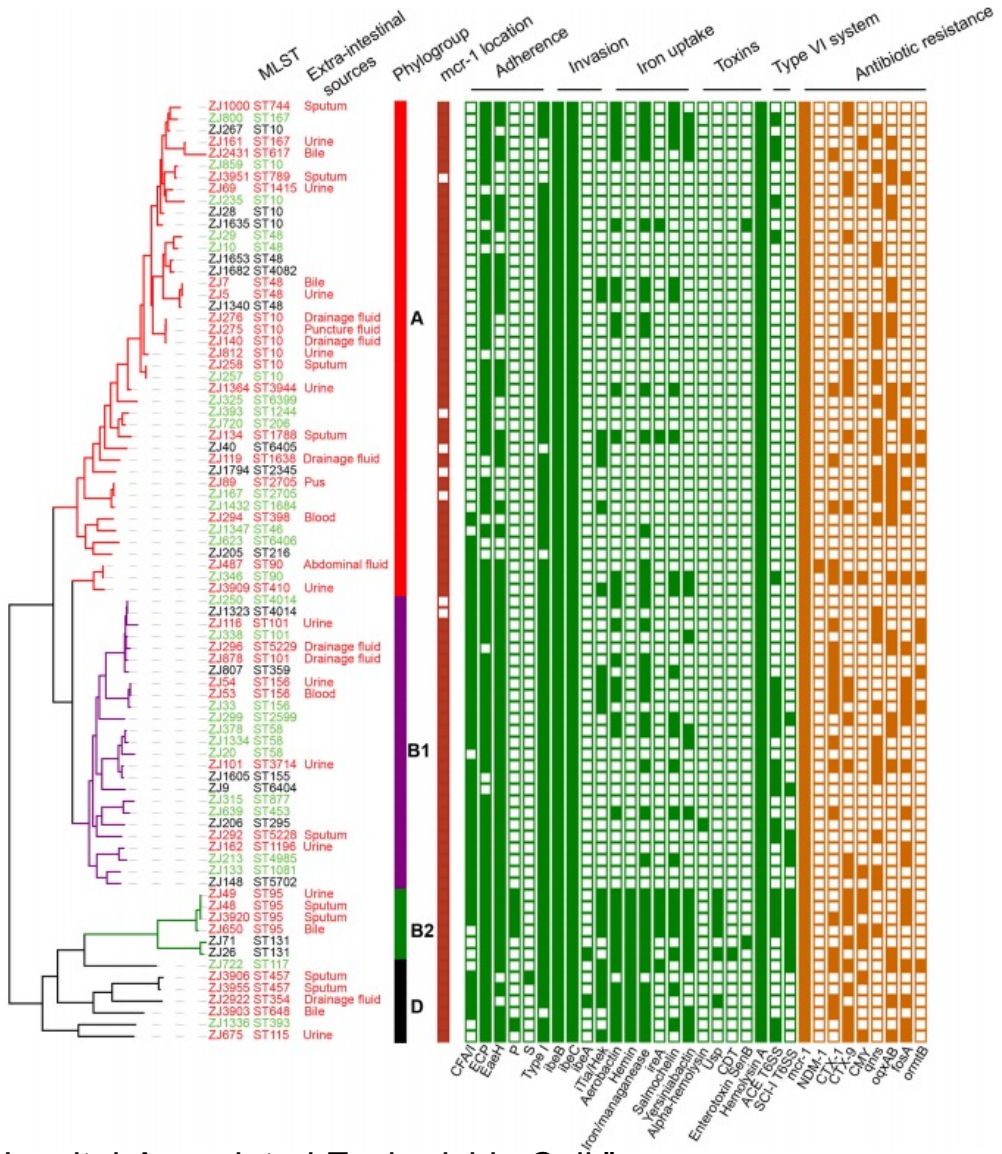
Genomic Sequence Analysis of Avian Flu



- Influenza A viruses caused 4 human pandemics in the past 100 years – new hypervirulent strains arise from mixing of human and animal flu viruses
- Birds are natural reservoir to influenza viruses

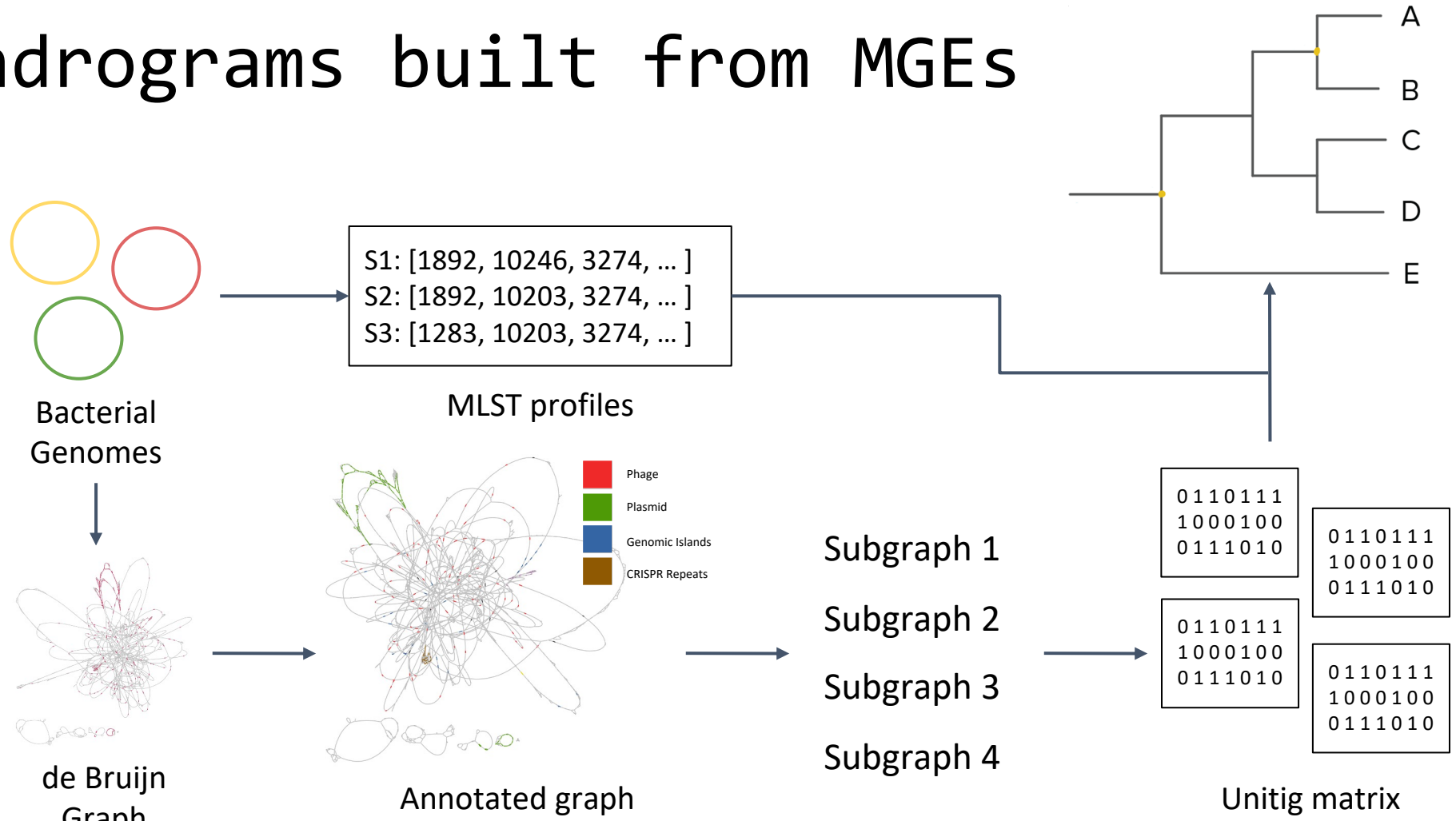
AMR and Mobile Element:

- Many antimicrobial resistant (AMR) genes can move around hosts (via mobile genetic elements)
- Detection of the genes by PCR or identification of the organisms alone are insufficient to understand the transmission of these AMR genes



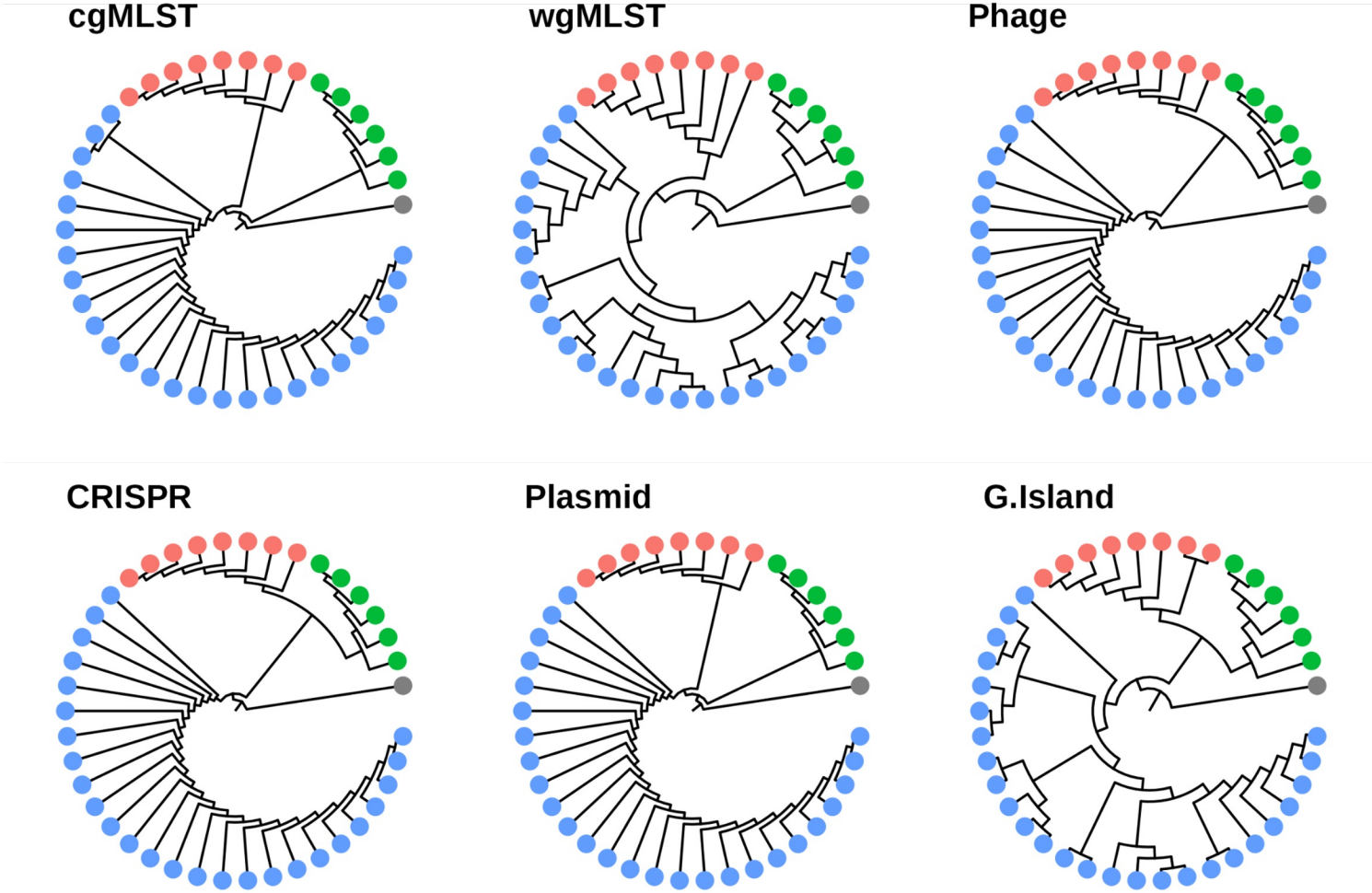
Shen et al. 2018. "Heterogeneous and Flexible Transmission of Mcr-1 in Hospital-Associated Escherichia Coli." https://doi.org/10.1128/mBio.00943-18.

Dendrograms built from MGEs



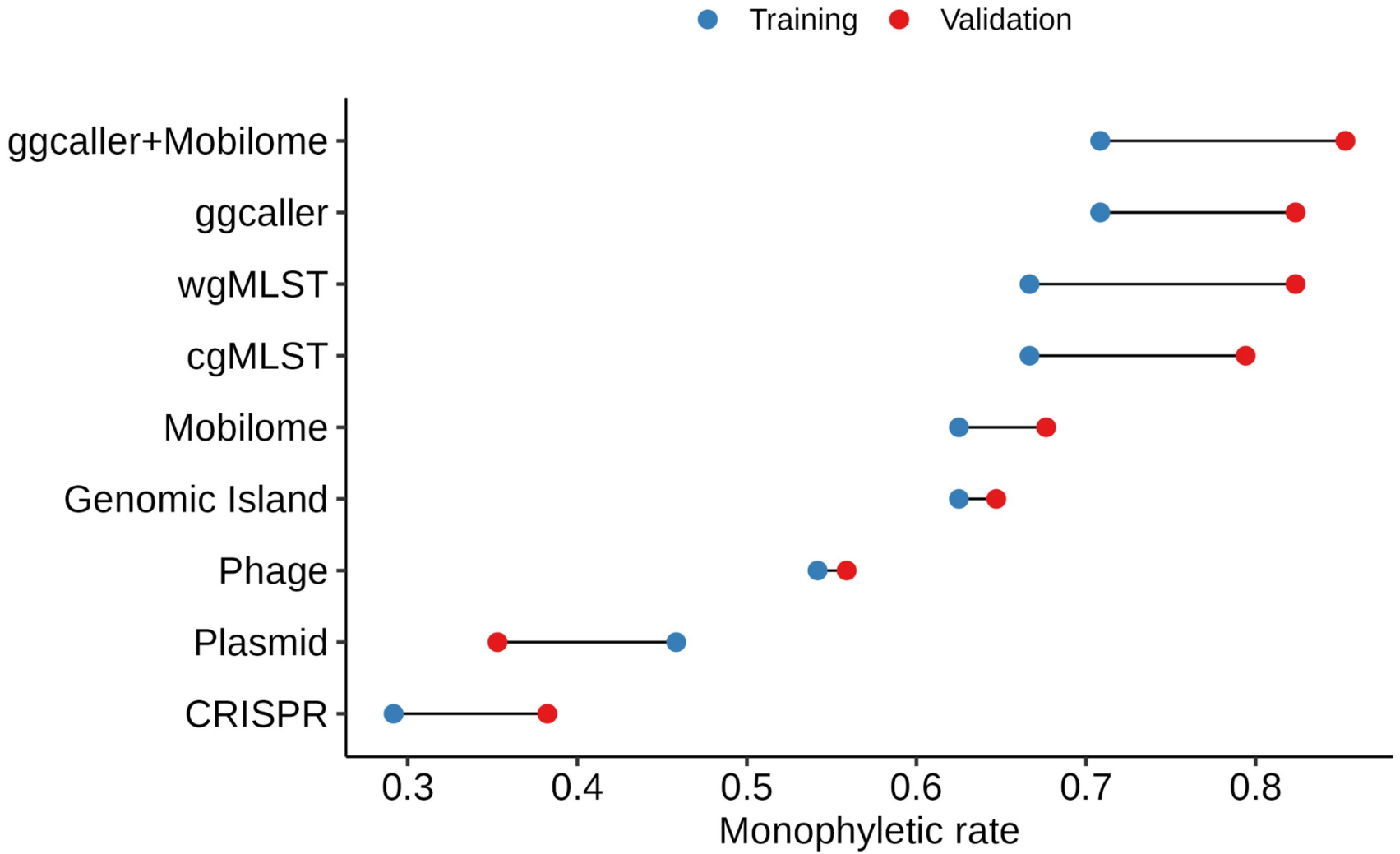
Liu and Hsiao, submitted

Topological comparison of dendrograms constructed from MGEs to systematics methods preferred by PulseNet



3 separate Salmonella outbreaks

Benchmarking performance of MGE clustering and gold standard subtyping methods



Genomics has been a hero of the COVID-19 pandemic

A SARS-CoV-2 vaccine candidate would likely match all currently circulating variants

Bethany Dearlove, Eric Lewitus, Hongjun Bai, Yifan Li, Daniel B. Reeves, M. Gordon Joyce, Paul T. Scott, Mihret F. Amare, Sandhya Vasan, Nelson L. Michael, Kayvon Modjarrad, and Morgane Rolland

PNAS September 22, 2020 117 (38) 23652-23662; first published August 31, 2020;
https://doi.org/10.1073/pnas.2008281117

The proximal origin of SARS-CoV-2

Kristian G. Andersen, Andrew Rambaut, W. Ian Lipkin, Edward C. Holmes & Robert F. Garry

Nature Medicine **26**, 450–452(2020) | Cite this article

5.03m Accesses | **706** Citations | **35003** Altmetric | Metrics

To the Editor – Since the first reports of novel pneumonia (COVID-19) in Wuhan, Hubei province, China[1,2], there has been considerable discussion on the origin of the causative virus, SARS-CoV-2[3] (also referred to as HCoV-19)[4]. Infections with SARS-CoV-2 are now widespread, and as of 11 March 2020, 121,564 cases have been confirmed in more than 110 countries, with 4,373 deaths[5].

SARS-CoV-2 is the seventh coronavirus known to infect humans; SARS-CoV, MERS-CoV and SARS-CoV-2 can cause severe disease, whereas HKU1, NL63, OC43 and 229E are associated with mild symptoms[6]. Here we review what can be deduced about the origin of SARS-CoV-2 from comparative analysis of genomic data. We offer a perspective on the notable features of the SARS-CoV-2 genome and discuss scenarios by which they could have arisen. Our analyses clearly show that SARS-CoV-2 is not a laboratory construct or a purposefully manipulated virus.

Cite as: X. Deng *et al.*, *Science* 10.1126/science.abb9263 (2020).

Genomic surveillance reveals multiple introductions of SARS-CoV-2 into Northern California

Xianding Deng[1,2]*, Wei Gu[1,2]*, Scot Federman[1,2]*, Louis du Plessis[3]*, Oliver G. Pybus[3], Nuno Faria[3], Candace Wang[1,2], Guixia Yu[1,2], Brian Bushnell[4], Chao-Yang Pan[5], Hugo Guevara[5], Alicia Sotomayor-Gonzalez[1,2], Kelsey Zorn[6], Allan Gopez[1], Venice Servellita[1], Elaine Hsu[1], Steve Miller[1], Trevor Bedford[7,8], Alexander L. Greninger[7,9], Pavitra Roychoudhury[7,9], Lea M. Starita[8,10], Michael Famulare[11], Helen Y. Chu[8,12], Jay Shendure[8,9,13], Keith R. Jerome[7,9], Catie Anderson[14], Karthik Gangavarapu[14], Mark Zeller[14], Emily Spencer[14], Kristian G. Andersen[14], Duncan MacCannell[15], Clinton R. Paden[15], Yan Li[15], Jing Zhang[15], Suxiang Tong[15], Gregory Armstrong[15], Scott Morrow[16], Matthew Willis[17], Bela T. Matyas[18], Sundari Mase[19], Olivia Kasirye[20], Maggie Park[21], Godfred Masinde[22], Curtis Chan[22], Alexander T. Yu[5], Shua J. Chai[5,15], Elsa Villarino[23], Brandon Bonin[23], Debra A. Wadford[5], Charles Y. Chiu[1,2,24]†

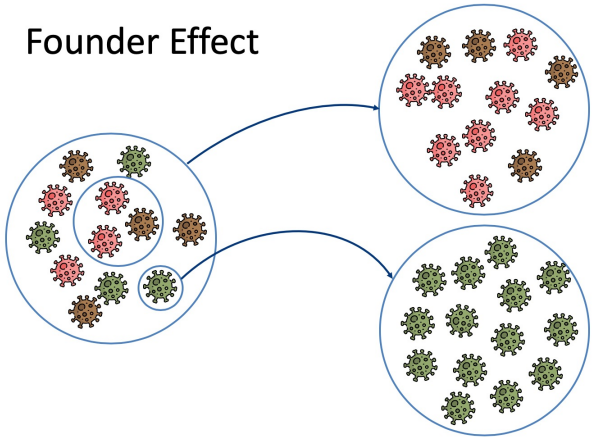
Comment on this paper

Large scale sequencing of SARS-CoV-2 genomes from one region allows detailed epidemiology and enables local outbreak management

Andrew J Page, Alison E Mather, Thanh Le Viet, Emma J Meader, Nabil-Fareed J Alikhan, Gemma L Kay, Leonardo de Oliveira Martins, Alp Aydin, David J Baker, Alexander J. Trotter, Steven Rudder, Ana P Tedim, Anastasia Kolyva, Rachael Stanley, Maria Diaz, Will Potter, Claire Stuart, Lizzie Meadows, Andrew Bell, Ana Victoria Gutierrez, Nicholas M Thomson, Evelien M Adriaenssens, Tracey Swingler, Rachel AJ Gilroy, Luke Griffith, Dheeraj K Sethi, Rose K Davidson, Robert A Kingsley, Luke Bedford, Lindsay J Coupland, Ian G Charles, Ngozi Elumogo, John Wain, Reenesh Prakash, Mark A Webber, SJ Louise Smith, Meera Chand, Samir Dervisevic, Justin O'Grady, The COVID-19 Genomics UK (COG-UK) consortium

doi: https://doi.org/10.1101/2020.09.28.20201475

Why sequence SARS-CoV-2 genomes



Transmission tracking at the regional, provincial, national and international

Cluster investigations (i.e. genomic epidemiology)

Evolving viral characteristics that might impact
- detection methods (PCR, serology)
- clinical outcomes (strain severity) and transmission
- The most reliable way to detect variants of concern
- effectiveness of healthcare measures, treatments and vaccines
  (ID regions of sequence frequently changing – or not changing at all – informing drug target/vaccine development and drug/vaccine continued effectiveness)

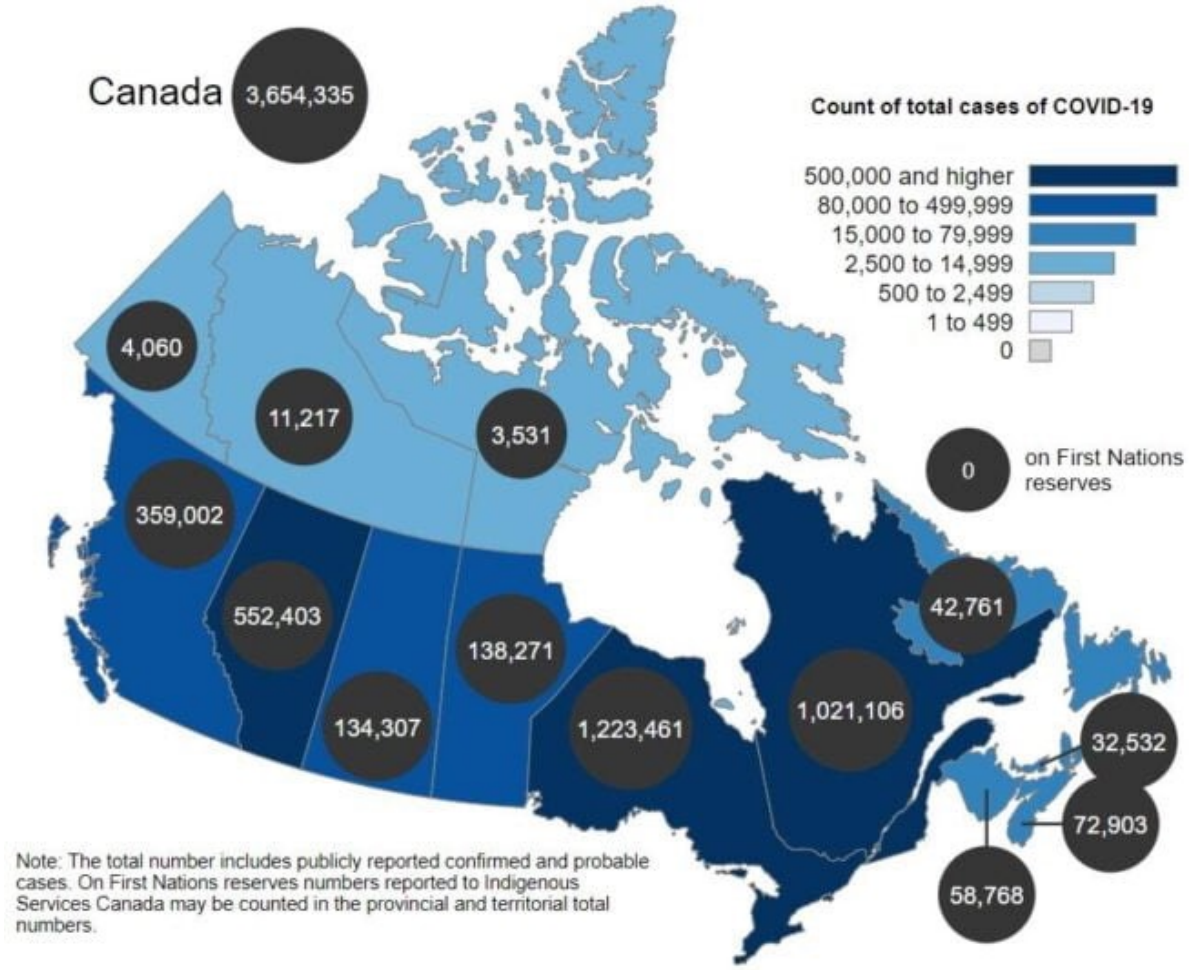
Many national and continental efforts to sequence the virus in close-to-real-time (COG-UK, SPHERES, Australia, Pan-Africa, etc) – however some of these efforts are now being sunsetted

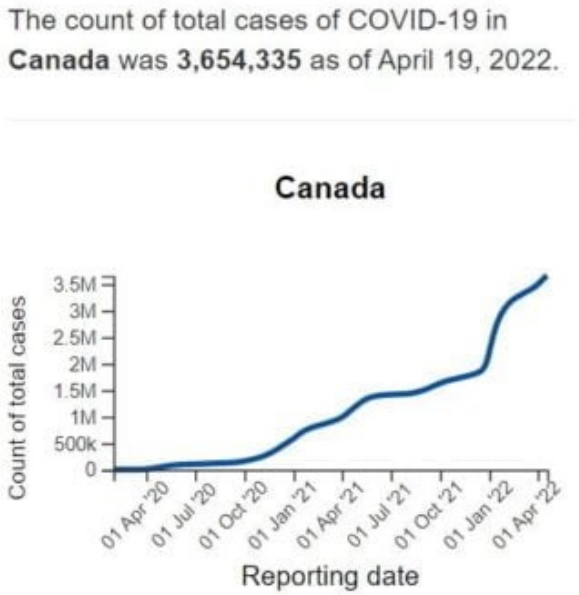
Canadian COVID-19 Genomics Network

- Established in March 2020 with an initial $40 Million Canadian Federal Government Investment
- $20 million viral genomic sequencing and genomic capacity building in public health
- $20 million human host genome sequencing from infected individuals
- Consortium of national and provincial public health laboratories, hospitals, research institutes, large-scale genome sequencing centers, industry, and coordinating centers.
- **Goals**
  - Coordinate and fund SARS-CoV-2 and human host genome sequencing efforts (up to 150,000 viral; 10,000 human)
  - Integrate **sequence data** and harmonize associated **clinical/epidemiological data (metadata)** – led by my group
  - Facilitate **data sharing** nationally and internationally
  - Capacity building, including for **future outbreak/pandemic preparedness**

The Canadian COVID-19 Genomics Network

- 150K viral genomes, 10K human genomes





To date: >500,000 viral genomes were sequenced in Canada

VirusSeq Data Portal – publicly accessible viral genomics data from Canada

Impact on Canadians

Genomic-based tracking and analysis of the evolving traits of the SARS-CoV-2 virus across Canada provides critical information for:

- Public health and policy decisions
- Testing and tracing strategies
- Virus detection and surveillance methods
- Vaccine development and effectiveness
- Drug discovery and effectiveness of treatment
- Understanding susceptibility, disease severity and clinical outcomes

Why Sequence this Virus?

Identify and track transmission trends at the regional, provincial, national and international scales.

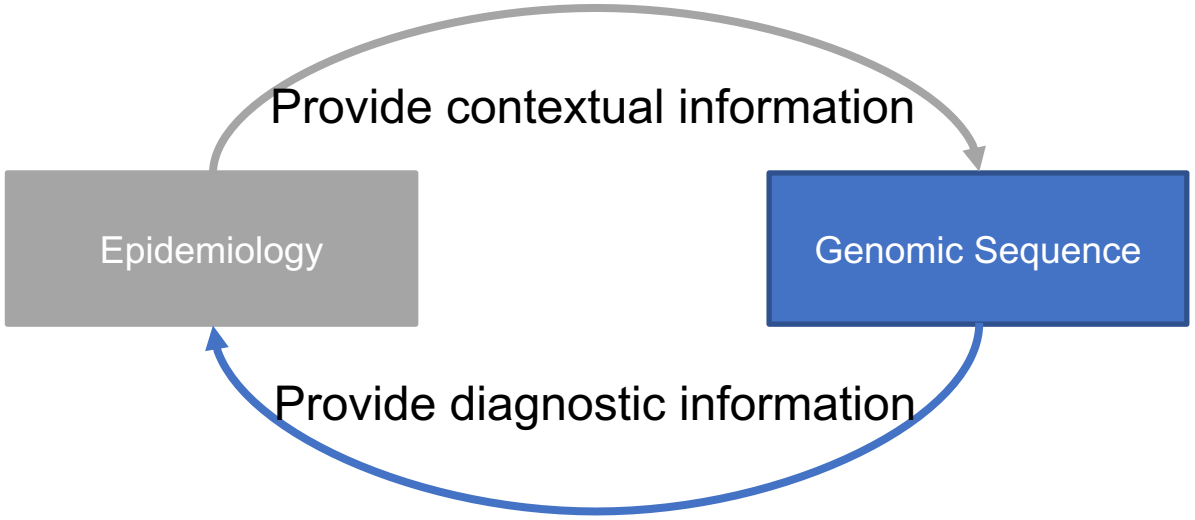
Aid detection of new clusters of cases/outbreaks.

Discover evolving viral characteristics that might impact.

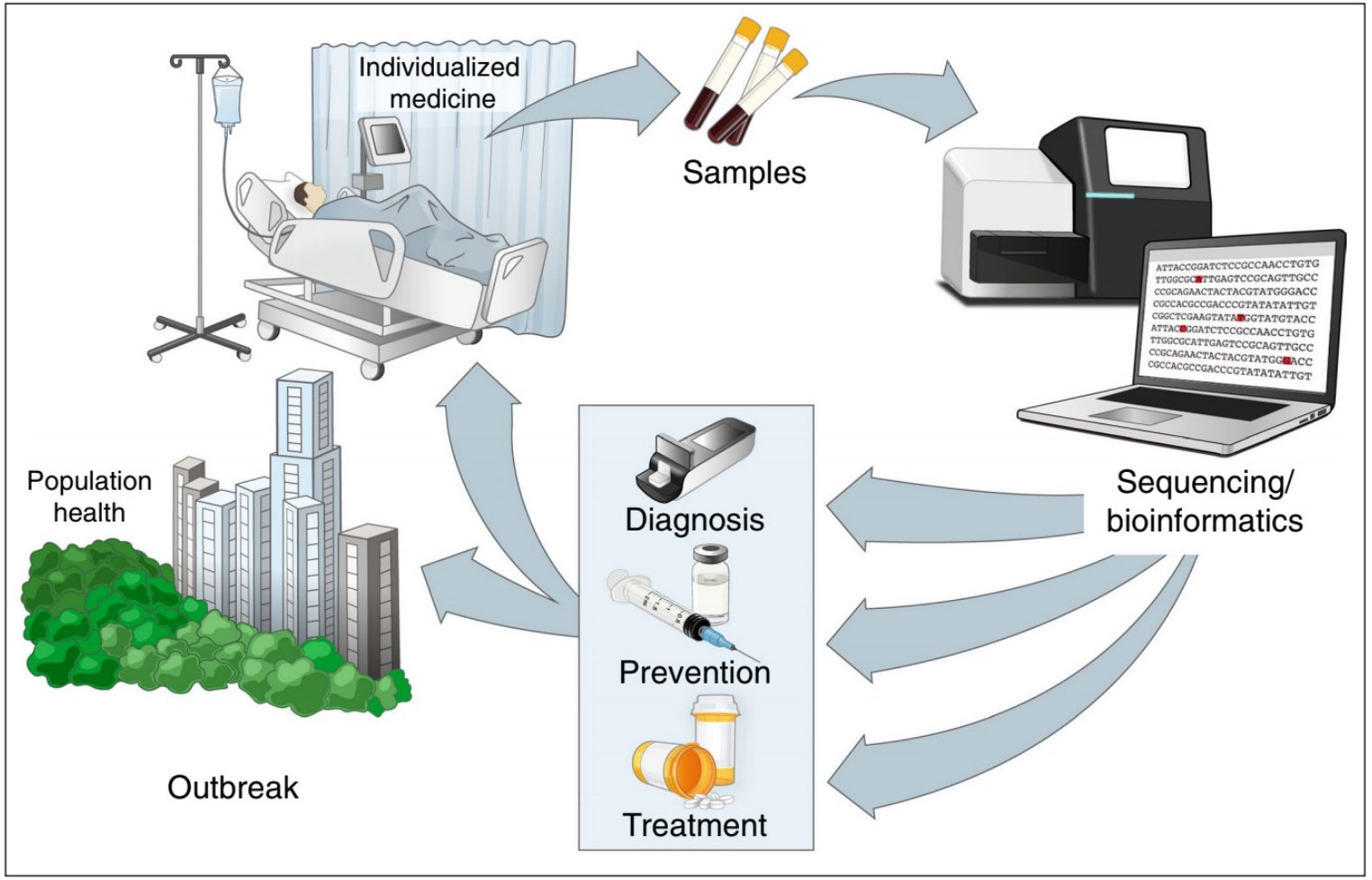
[2405.04734] The Canadian VirusSeq Data Portal & Duotang: open resources for SARS-CoV-2 viral sequences and genomic epidemiology (arxiv.org)

Genomic Epidemiology

- Def: Combine **whole genome sequencing** data from **pathogens** with **epidemiological investigations** to track the spread of an infectious disease



Whole Genome Sequencing Based Workflow



Ladner et al. 2019 Nature Med

Benefits and Challenges

- Simplified Workflow
- Faster turn-around-time (for some applications)
- Cost-saving by reducing the number of platforms/instruments
- Sequencing is becoming commoditized
- Results (sequences) more **comparable** and **sharable** than other test data
- Value-added analysis (e.g. pathogen evolution, AMR prediction, transmission dynamic modeling etc.)

- Results are harder to process and interpret
- Computational Resource Requirement higher (no IT support?)
- Rapid changing technologies
- Per sample cost still higher? Batching required
- Other benefits/challenges?

High Throughput Sequencing (HTS)

- HTS = next-gen sequencing and third-gen sequencing platforms
- Sequence data have many clinical and PH lab utilities
    - Diagnostic – strain level identification, virulence gene ID, AMR gene ID
    - Surveillance –gene-by-gene typing, single-nucleotide-variant (SNV) typing, serotyping, copy-number variants (VNTR, MLVA)
    - Outbreak detection and investigations
    - Source Tracking
- These genomic data can be useful for downstream research use (e.g. improve our understanding of pathogen evolution)

Illumina MiSeq (NGS)

Oxford MinION (3rd Gen)

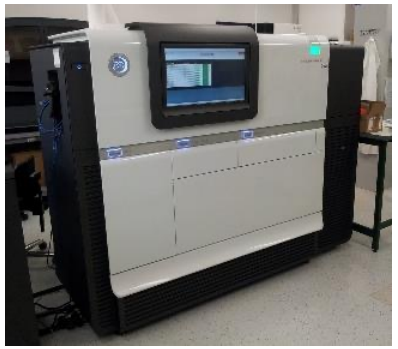
Sequencing technologies

Sanger

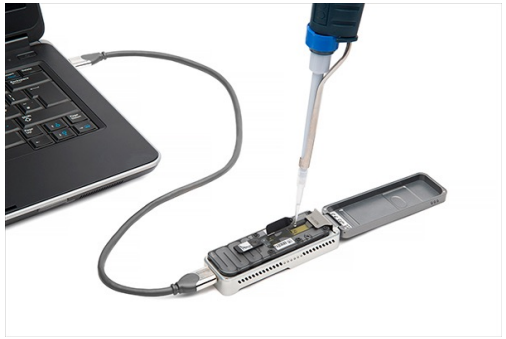
Gene Studio
(Ion Torrent)

Roche 454

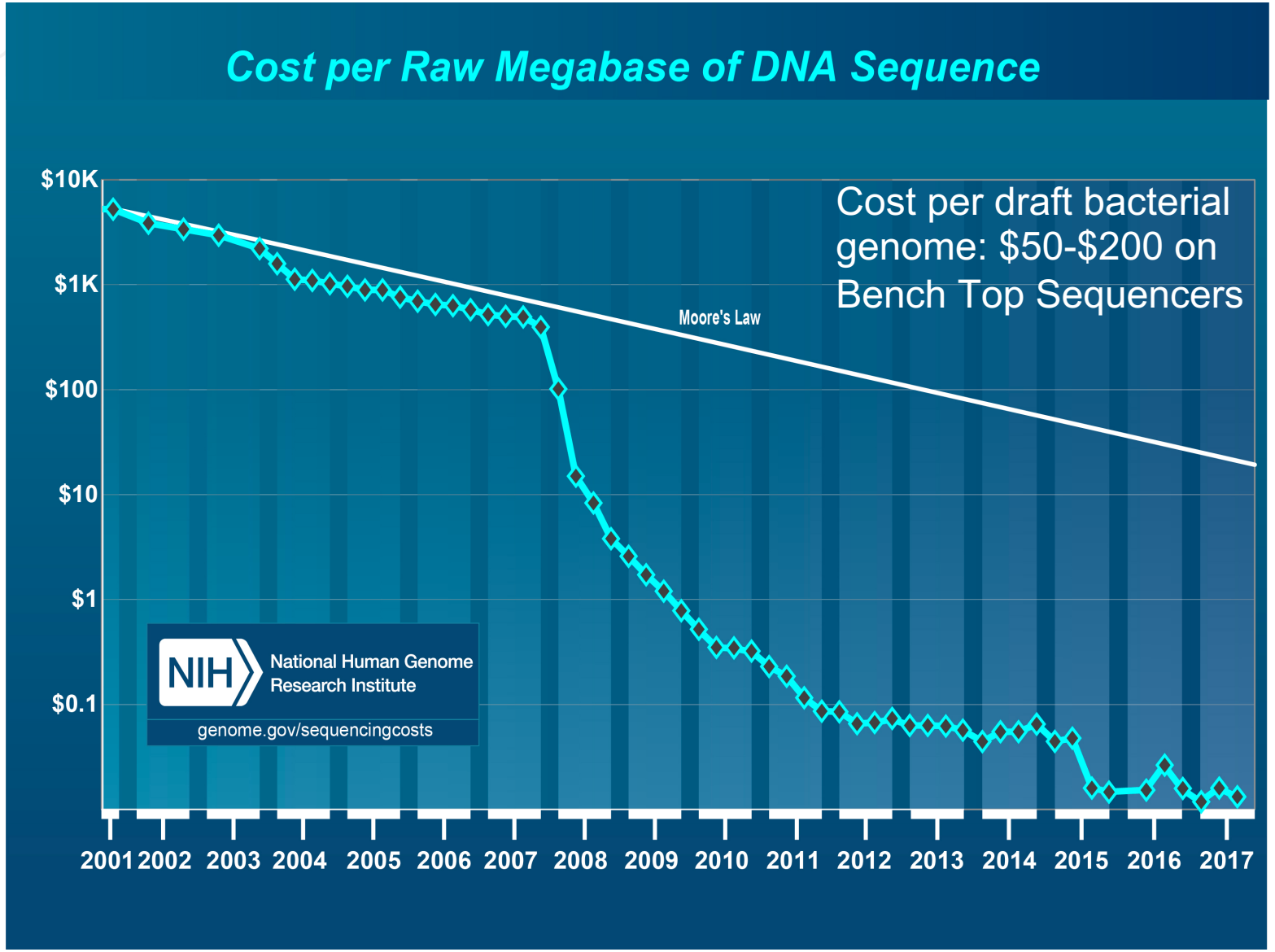
Illumina *Seq



Pacific Biosciences



Nanopore



Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata.

Short vs. Long Read Sequencing

Sequencing technology (specifications)	Instrument cost	Library prep cost	Cost per lane/cell	Cost per million reads	Cost per billion bases	Runtime (h)	Throughput (k reads at maximum runtime)	Avg read length (maximum)	Read accuracy (%)	Error profile
PacBio Sequel I (maximum 8 cells)	300,000	150–400	1,000	2,000	100	10–20	500	20,000 (50,000)	85–88/99.9 (CCS)	Homopolymers
PacBio Sequel II (maximum 16 cells)	650,000	75–400	2,000	500	17	10–30	4,000	30,000 (100,000)	88–90/99.9 (CCS)	Homopolymers
ONT Flongle	1,300	40–90	80	400	13	0.1–12	200	30,000 (60,000)	96–99	Homopolymers
ONT MinION	900	90–130	600	600	12	0.1–48	1,000	50,000 (2.3M)	96–99	Homopolymers
ONT GridION (maximum 5 FCs)	45,000	90–130	600	600	12	48	1,000	50,000 (2.3M)	96–99	Homopolymers
ONT PromethION (maximum 48 FCs)	176,000	90–600	1,400	250	8	72	6,000	30,000 (330,000)	96–99	Homopolymers
SLR (example of NovaSeq 2 × 150 PE, 30× coverage, for 5 kb)	900,000	30–50	5,000	1,250	140	44	4,000	9,000 (12,000)	100	Unequal coverage
Illumina NovaSeq S4 (2 × 150 PE; maximum 4 lanes)	900,000	50–100	5,000	2.5	9	44	2,000,000	250 (290)	99.9	Low-quality ends
Illumina MiSeq (2 × 300 PE; maximum 2 lanes)	100,000	50–100	2,000	100	175	56	20,000	550 (590)	99.9	Low-quality ends
ION G5-S5 Prime, P550 chip (maximum 2 chips)	180,000	50	700	5	25	6.5	130,000	200 (250)	99.0–99.5	Homopolymers
ION G5-S5, P530 chip 600 SE	60,000	50	500	40	70	7	12,000	570 (650)	99.3–99.7	Homopolymers, low-quality ends
MGI Tech DNBSEQ-T7 (2 × 150 PE; maximum 4 cells)	990,000	50–100	6,000	1.2	4.5	24	5,000,000	250 (290)	99.9	Low-quality ends
MGI Tech DNBSEQ-G400RS (2 × 200 PE; 400 SE maximum 2 cells × 4 lanes)	480,000	50–100	1,800	4	11	108	450,000	350 (400)	99.9	Low-quality ends

a Costs are given in EUR. Information is compiled from multiple recent literature sources and price requests from sequencing companies. FC, flow cell; PE, paired-end; SE, single-end.

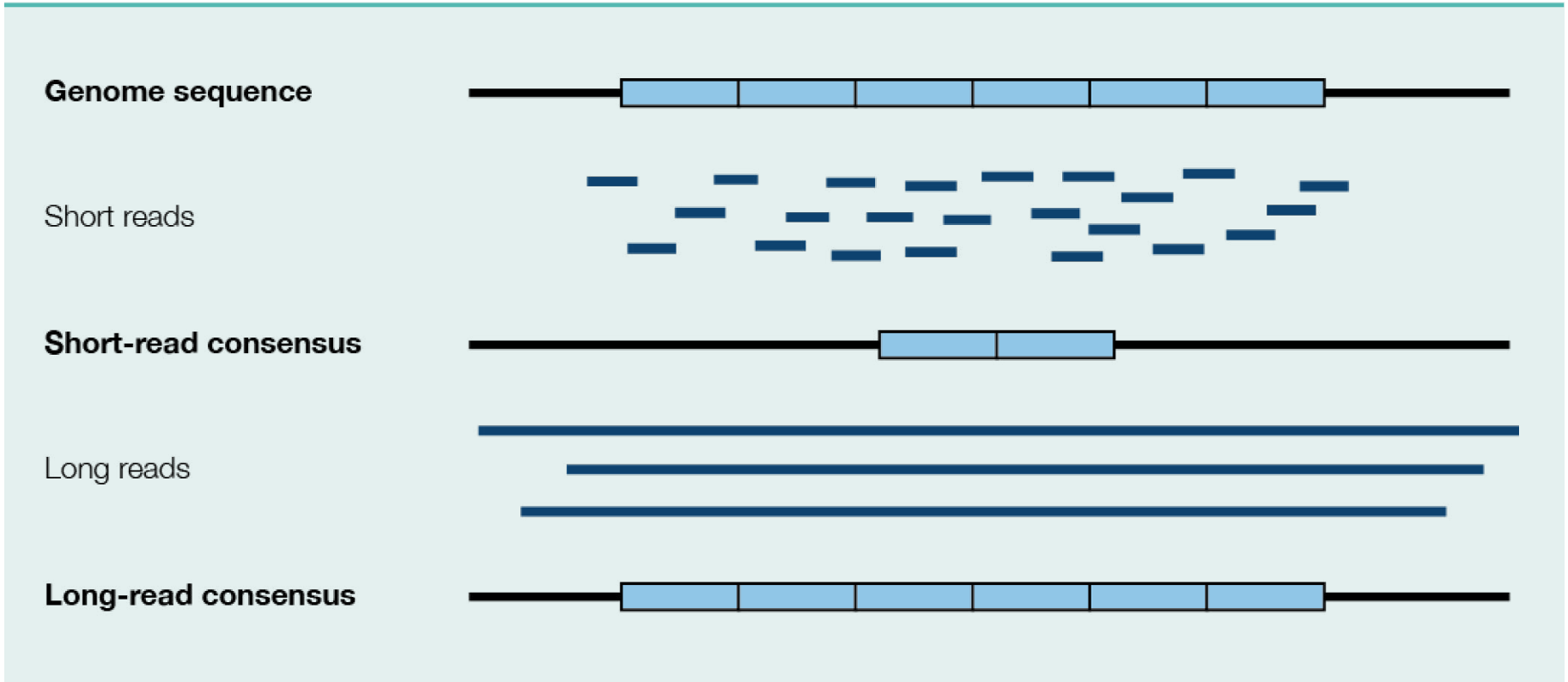
Tedersoo L et al Perspectives and Benefits of High-Throughput Long-Read Sequencing in Microbial Ecology. Appl Environ Microbiol. 2021;87: e0062621. doi:10.1128/AEM.00626-21

Short vs. Long Read Sequencing Summary

- Short Read:
  - Cheaper (per base)
  - Higher capacity / throughput
  - Higher accuracy
  - Reads are consensus of many molecules

- Long Read:
  - More expensive (per base)
  - Lower capacity /throughput
  - Lower accuracy
  - Capable of single molecule sequencing

Why are long reads beneficial?



- *de novo* assembly of short read sequences – repetitive regions

ONT website

NGS Error Rates

- Sequencing instruments have different error rates and are prone to different error types
  - **Sanger** – prone to **substitution** errors and 0.1-1% error rate
  - **Ion Torrent** – prone to **indel** and 1% error rate
  - **SOLiD** – prone to **A-T bias** and >0.06% error rate
  - **Illumina** – prone to **substitution** errors and >0.1% error rate
  - **PacBio** and **MinION (3rd Gen Sequencer)** >5% error rate

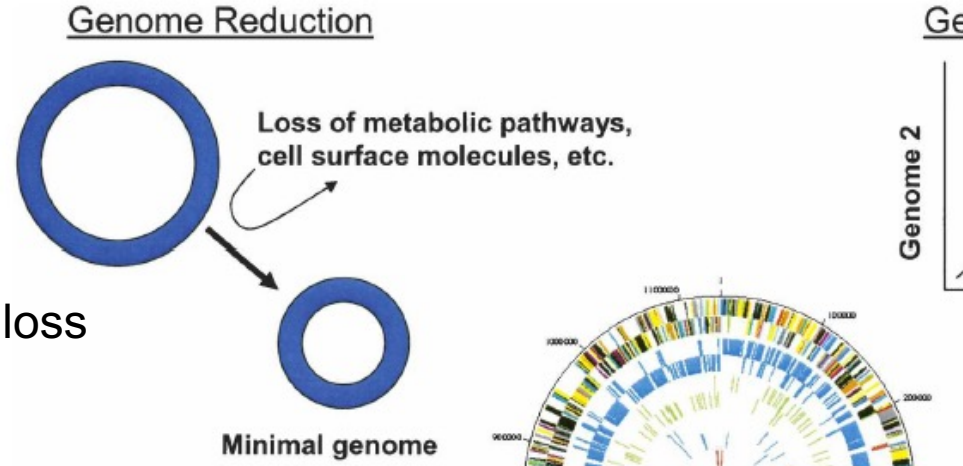
- To minimize errors, the same regions of the genome are typically sequenced multiple times (>30X). This is called sequencing **depth coverage.** The consensus is taken as the correct sequence.

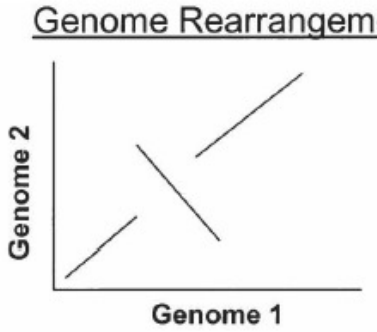
Pathogen Genomes

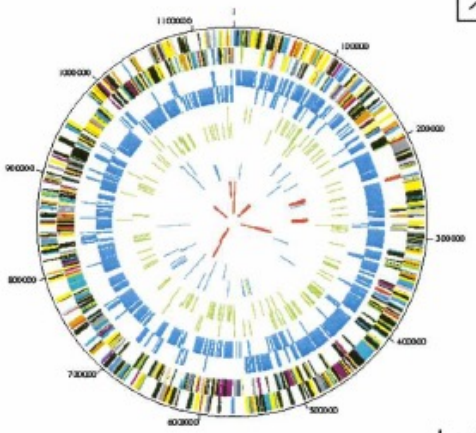
- Bacteria:
  - Typically contained within a **single** circular chromosome (some are linear)
  - **Haploid** genomes
  - May contain **plasmids** (extrachromosomal DNA)
  - Genome size range from 0.5Mb to ~10Mb (average is about 3-5Mb and contain about 3000-5000 genes)
- Viruses:
  - Can be DNA or RNA, single stranded or double stranded (classified into 7 families)
  - Range from 1-2Kb to ~1-2Mb (Pandoravirus salinus = 2.5Mb!)
  - Depend on host cellular mechanisms to replicate
- Eukaryotic Parasites (fungi, protists, and worms):
  - Usually a few to a few hundred Mbs
  - Usually multiple chromosomes

Microbial Genomes are Constantly Evolving

Specialization
“lean and mean”
via gene deletion / loss

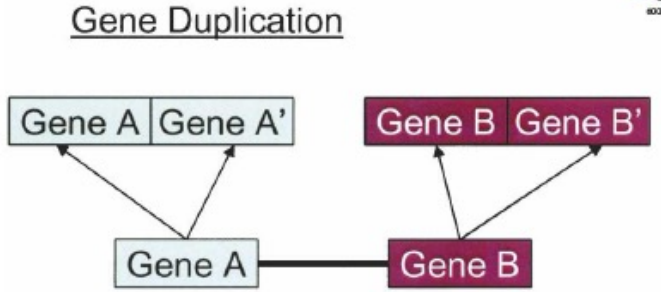






Gene expression
and be turned on
and off

New
function can
be derived
through:



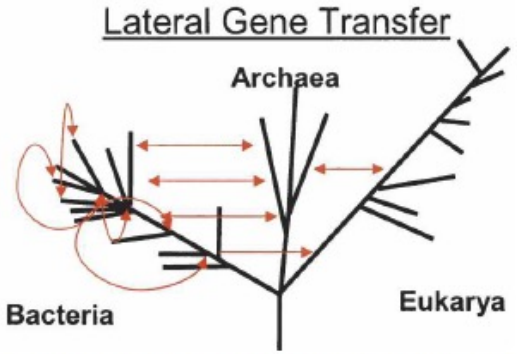
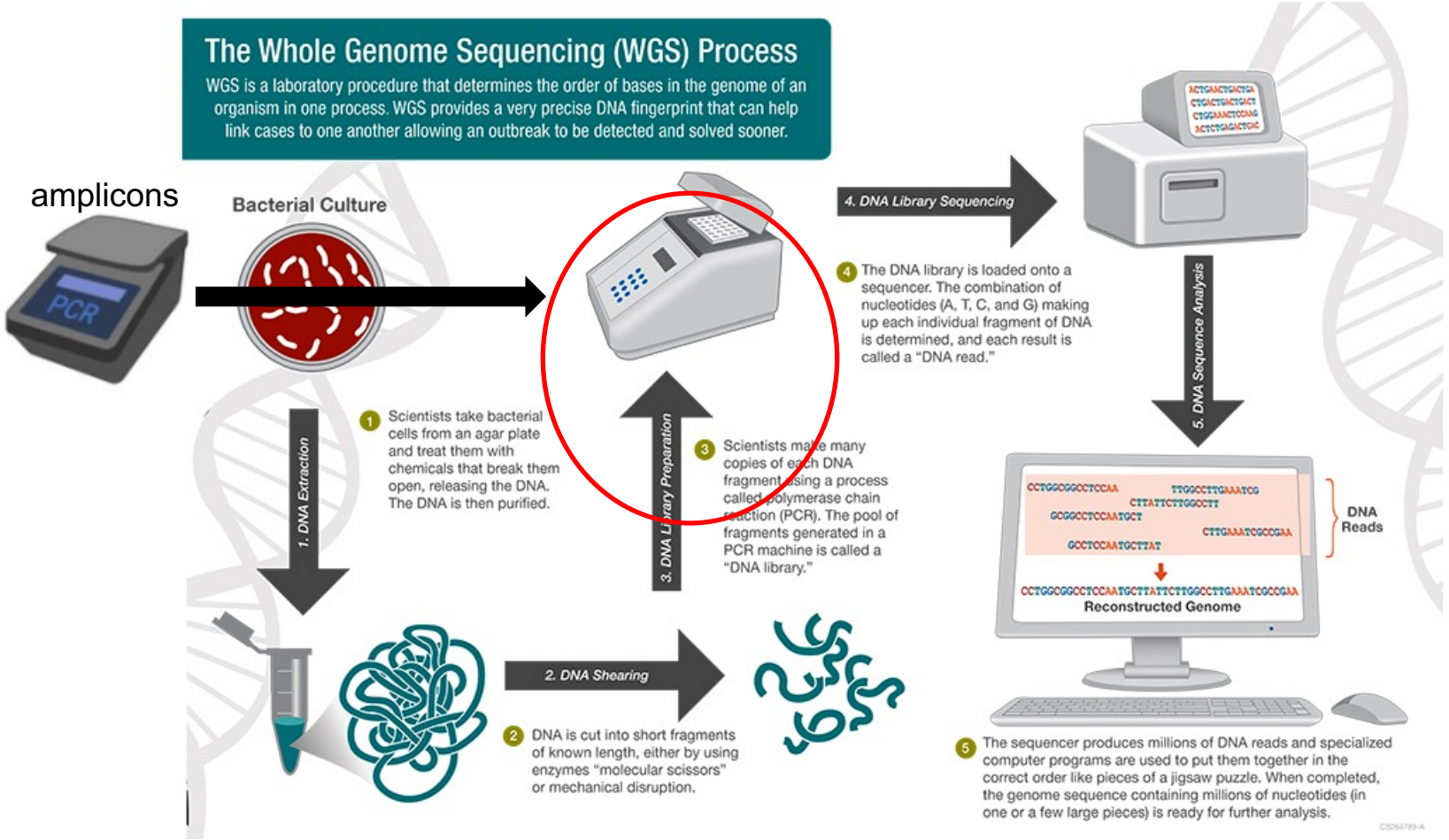


Figure 2. Multiple forces, including genome reduction, genome rearrangement, gene duplication, and acquisition of new genes via lateral gene transfer, are shaping microbial genomes. Details of each of these processes can be found in the text.

Whole Genome Shotgun Sequencing with NGS



https://www.cdc.gov/pulsenet/pathogens/protocol-images.html#wgs

Sequence Data Analysis

- Millions to Billions of partially overlapping reads were generated from a single “sequencing run”
- Steps of Data Analysis
  - Assembly
    - De-novo Assembly
    - Mapping
  - Annotation (adding information to sequences)
    - Gene Prediction (determine coding vs. non-coding regions of the genome)
    - Functional Prediction (determine the functions of genetic elements)
  - Variant Identification (single nucleotide variants, Allelic differences)

Contigs vs. Complete Genomes

- After assemblies, there are often still “gaps” in the genomes that can not be closed due to lack of sequence coverage or, more likely, un-resolved repeats
- So instead of the complete genome, you get a set of contiguous sequences (**contigs**) representing most of the genome
- Closing the gaps manually (aka “**finishing**” the genome) is labor-intensive and expensive so very few groups still do that
- Long reads from 3rd generation sequencers can improve assembly

Genome Annotation

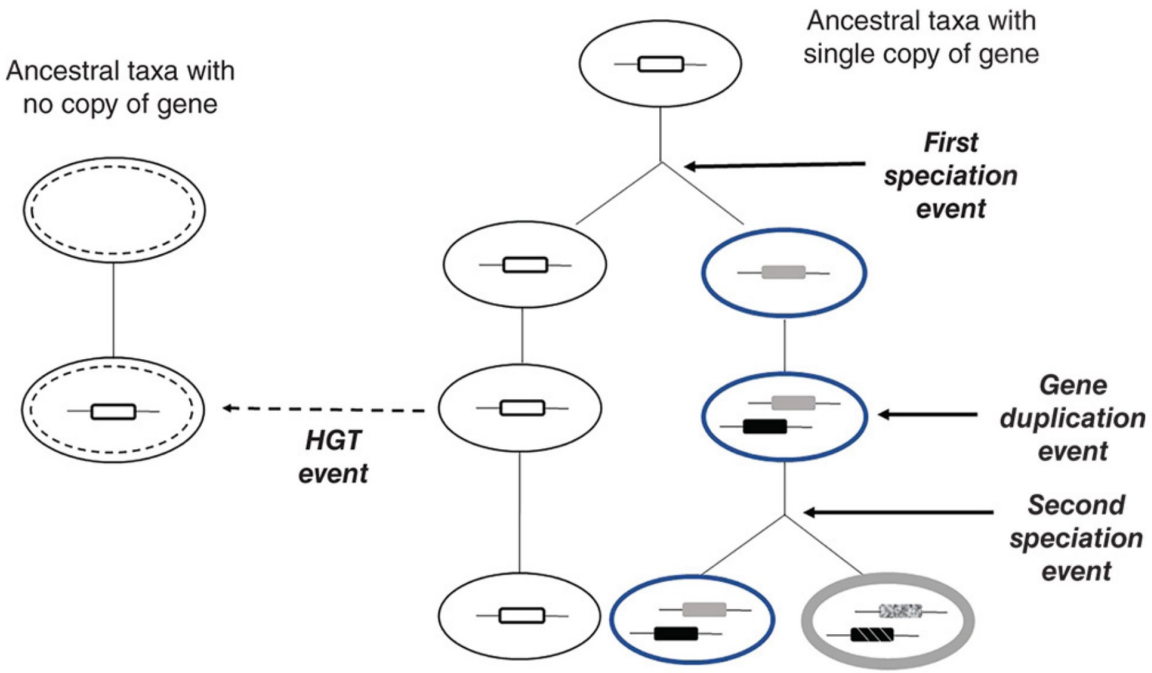
- Strings of A's, T's, C's and G's do not mean much on their own to us
- The goal of genome annotations is to identify **genes** and other features and locate them on the genomic sequence
  - Functions
  - Locations
- Locations of coding genes in a sequence can be identified by computer program because coding genes have different "word" frequencies compared to non-coding sequences – **ab initio gene prediction**
  - Works well for microbial genomes which are compact and usually no intron

Function Prediction

- The most common way to assign functions to a protein-coding gene is by **sequence similarity search**
- **It is assumed that genes that have sequence similarity are derived from the same ancestral gene and, therefore, have similar functions**
- BLAST is the most common tool for performing similarity search.
- Infer the function of one gene/protein based on its similarity to another gene/protein of known function is called "**transitive annotation**"
- Requires a database of known genes (e.g. GenBank)

Homology

- Homology: similarity due to **shared common ancestry**
- **Orthologs**: arise due to speciation
- **Paralogs**: arise due to gene duplication
- **Xenologs**: arise due to horizontal gene transfer (no homology!)



If the black gene is deleted in one sister taxon and the gray gene is deleted in the other; this can result in mis-interpretation of the true relationship of the two genes

HGT event could also lead to misinterpretation of phylogeny

BLAST

- Basic Local Alignment Search Tool
- Developed in 1990 and 1997 (S. Altschul)
- A heuristic method for performing local alignments through searches of high scoring segment pairs (HSP's)
- 1st to use statistics to predict significance of initial matches - saves on false leads
- Offers both sensitivity and speed
- Most highly cited bioinformatics tool!
- Many more sequence similarity search tools that are much faster and equally sensitive

Comparative Genomics

- The goal of comparative genomics is to identify **genomic variations** that can be correlated to phenotypic characteristics of an organism

- For example: we might be interested to know why certain isolates of a pathogen are more resistant to antibiotics or more virulent
  - E.g. comparing two strains of *E. coli* could reveal that one contains Shiga toxins (bloody diarrhea) and the other does not (commensal)

- We can also use these variations to track the transmission of pathogens

Comparative Genomics

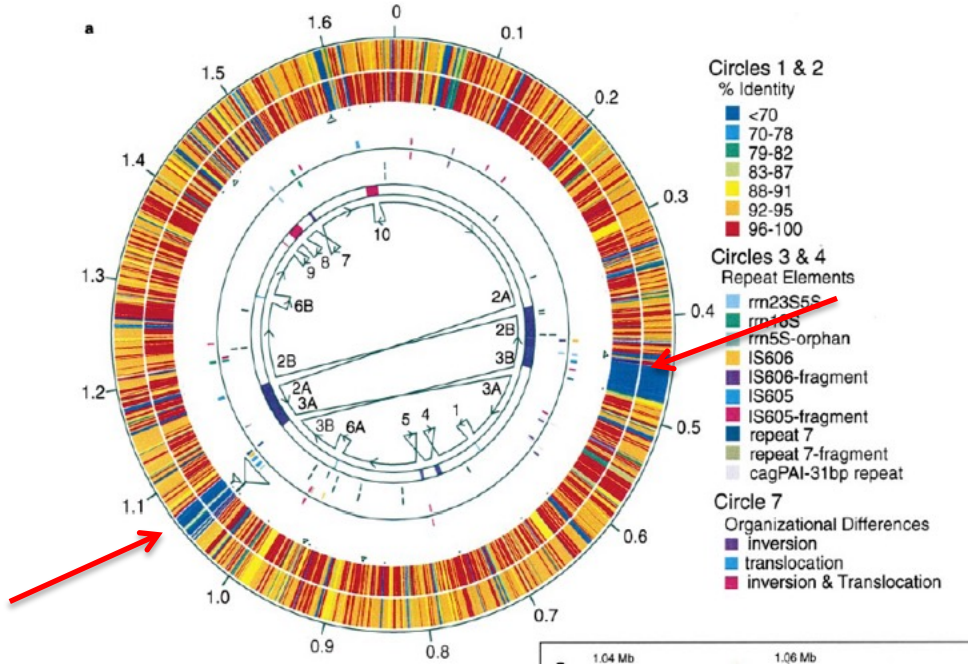
- Variations can occur at different levels
  - **Regional** (a stretch of DNA is present in one isolate but absent in another)
    - Strain-specific chromosomal regions
    - Strain-specific plasmids
    - VNTRs (tandem repeats)
  - **Gene** (a gene is missing or codes for different amino acids in one isolate compared to another)
    - Strain-specific genes
    - Allelic differences
    - Only looking at genes and not regulatory elements or non-coding genes
  - **Nucleic Acid** (a single nucleotide is different in one isolate compared to another)
    - Single nucleotide variants (SNVs) or single nucleotide polymorphism (SNPs)

First Comparative Genomics Paper

- published in 1999
  - 2 *Helicobacter pylori* genomes isolated 7 years apart were compared

Found more than half of the **strain specific genes** are clustered in hyper variable regions (red arrows)

This observation soon was consistently observed in many other species

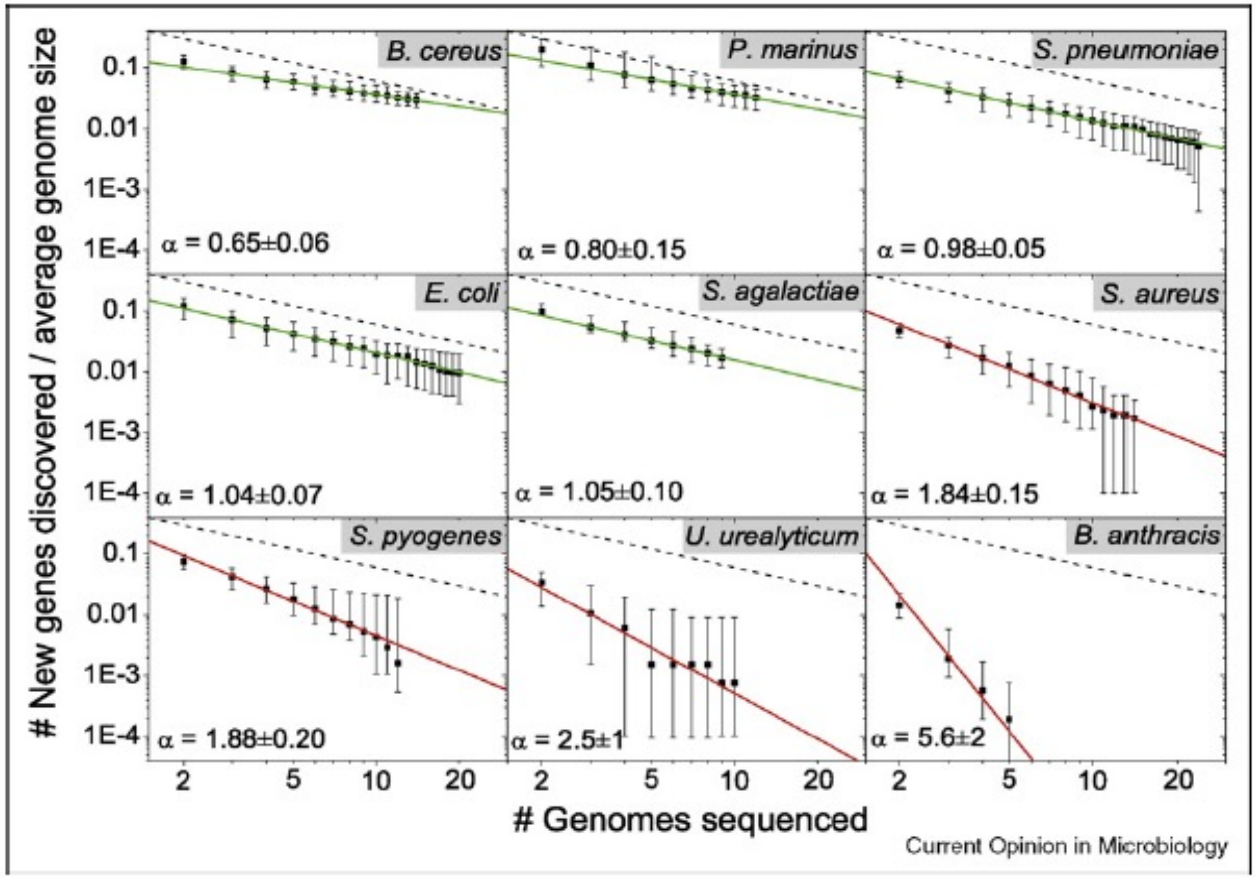


Alm et al, Nature 1999

Pan-genomes

- Comparative Genomics and huge genomic variations among strains in a bacterial species lead to the idea of pan-genomes
- The term first coined in 2005 in a paper by Tettelin et al., in which they compared sequenced genomes from six *S. agalactiae (group B Streptococcus)*.
- Pan-genome is the entire gene set of a species - consists of the **core (housekeeping)** genes of strains in a species + **strain-specific (accessory)** genes
- Pan-genome calculation extrapolates observations based on a limited number of strains to come up with the theoretical number of genomes required to fully capture the pan-genome of a species

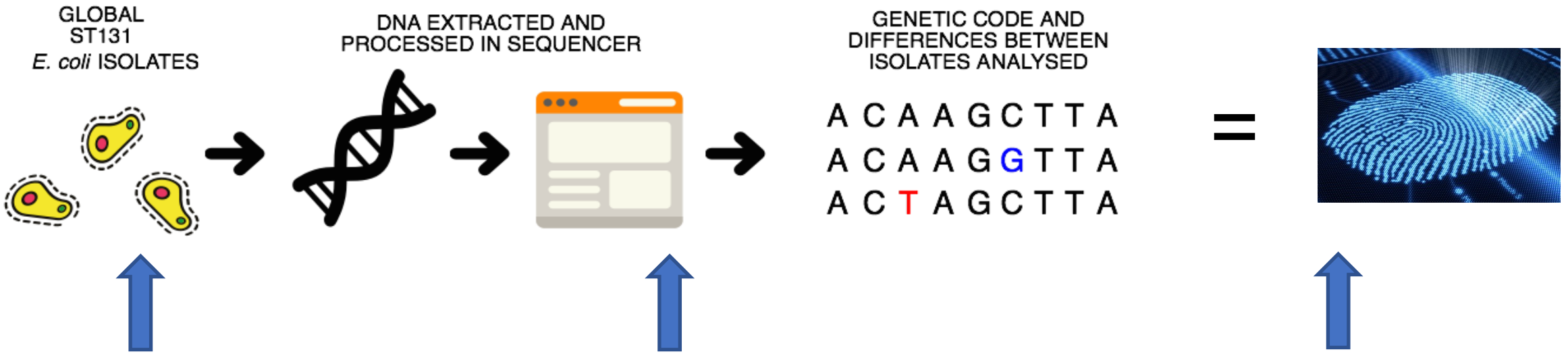
Open vs. Closed pan-genome



If the number of strains needed to capture the pangenome of a species is finite (red lines), then the species have a **closed pan-genome**; otherwise it has an **open pan-genome**.

This concept helps microbiologists predict how a species will evolve over time (e.g. is it likely to acquire many resistance genes?).

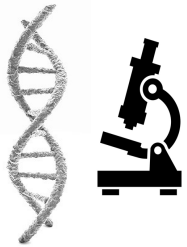
Context is everything!




Sample metadata


Clinical/Epi data

Methods


Lab results
(geno/pheno)

Significant Amount of Missing Contextual Data (labels)

Global landscape of SARS-CoV-2 genomic surveillance and data sharing

Zhiyuan Chen, Andrew S. Azman, Xinhua Chen, Junyi Zou, Yuyang Tian, Ruijia Sun, Xiangyanyu Xu, Yani Wu, Wanying Lu, Shijia Ge, Zeyao Zhao, Juan Yang, Daniel T. Leung, Daryl B. Domman & Hongjie Yu

Nature Genetics **54**, 499–507 (2022) | Cite this article

Moreover, incomplete metadata attached to GISAID sequences was common globally, with about **63% of sequences missing demographic information** (age and sex), **84% of sequences missing sampling strategy** and more than **95% of sequences missing patient-level clinical information**

Challenges with Data Sharing in Canada

- Canada is comprised of **14 distinct healthcare** systems
- N**o universal standard** for data collection or sharing
- No legally binding public health **data sharing agreement** in Canada
- Arguably, the restrictive flow of patient data across Canada is a violation of the **universality** and **portability** of healthcare

Attaran, A. & Houston, A. Pandemic Data Sharing: How the Canadian Constitution Turned Into a Suicide Pact. (2020). doi:10.2139/ssrn.3612825

frontiers in Genetics

POLICY AND PRACTICE REVIEWS
published: 24 March 2022
doi: 10.3389/fgene.2021.716541

Addressing Privacy Concerns in Sharing Viral Sequences and Minimum Contextual Data in a Public Repository During the COVID-19 Pandemic

Lingqiao Song[1†], Hanshi Liu[1†], Fiona S. L Brinkman[2], Erin Gill[2], Emma J. Griffiths[3], William W. L Hsiao[2], Sarah Savić-Kallesøe[2], Sandrine Moreira[4], Gary Van Domselaar[5], Ma'n H. Zawati[1] and Yann Joly[1]*

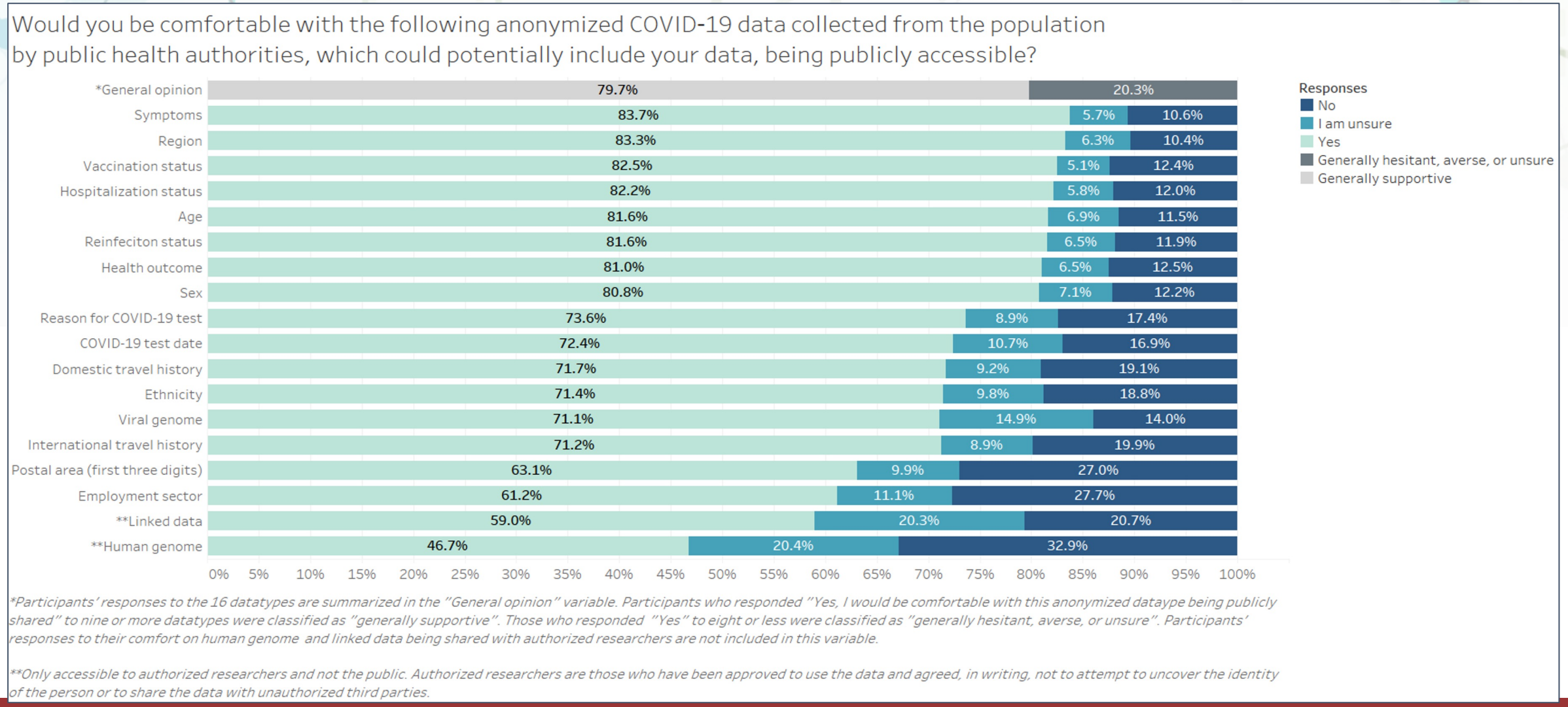
Open access Original research

BMJ Open

Canadians' opinions towards COVID-19 data-sharing: a national cross-sectional survey

Sarah A Savic Kallesoe ,[1,2] Tian Rabbani ,[1,3] Erin E Gill,[4] Fiona Brinkman,[4] Emma J Griffiths,[1] Ma'n Zawati,[5] Hanshi Liu,[5] Nicole Palmour,[5] Yann Joly ,[5] William W L Hsiao[1]

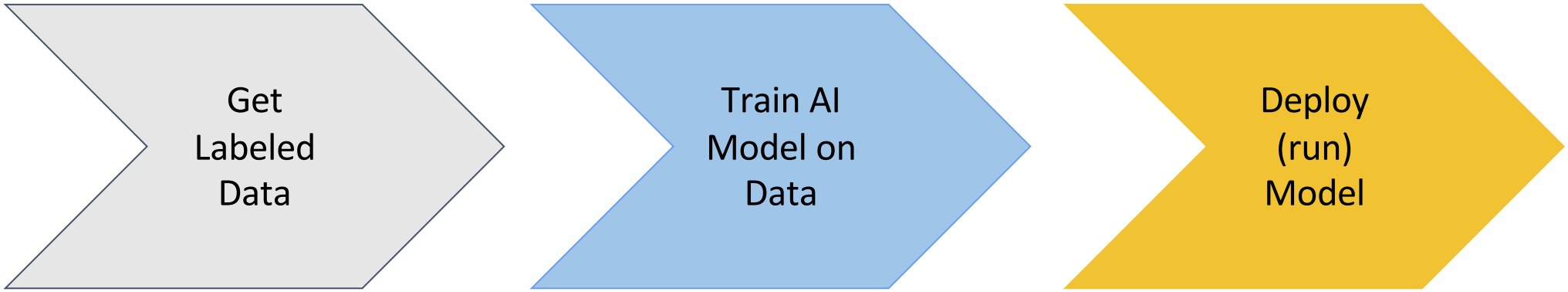
Canadians favour de-identified public health data sharing



**Participants' responses to the 16 datatypes are summarized in the "General opinion" variable. Participants who responded "Yes, I would be comfortable with this anonymized datatype being publicly shared" to nine or more datatypes were classified as "generally supportive". Those who responded "Yes" to eight or less were classified as "generally hesitant, averse, or unsure". Participants' responses to their comfort on human genome and linked data being shared with authorized researchers are not included in this variable.*

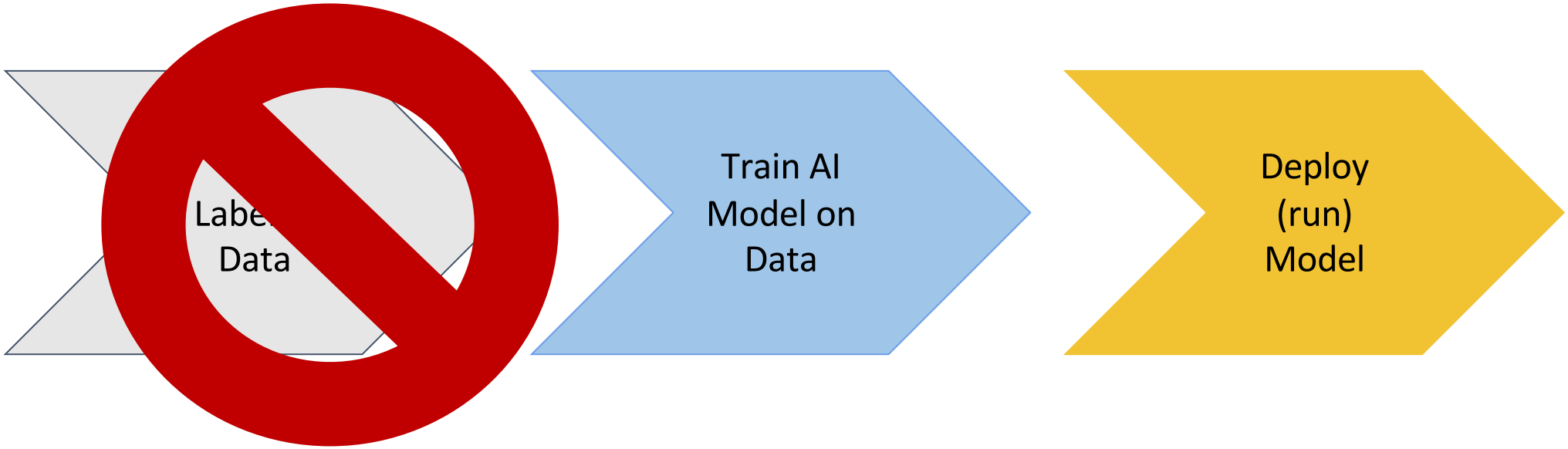
***Only accessible to authorized researchers and not the public. Authorized researchers are those who have been approved to use the data and agreed, in writing, not to attempt to uncover the identity of the person or to share the data with unauthorized third parties.*

General Process of Training an AI Model



See Andrew Ng's lecture: (65) Andrew Ng: Opportunities in AI - 2023 - YouTube

Getting Labeled Data is a Major Bottleneck



Take Home Messages

- Pathogen genomic sequence data provide valuable information for infectious disease epidemiological investigations
- While there is a wealth of microbial and viral genomic sequence data, the contextual information needed to make the data valuable for (re)analysis is severely lacking
- We need to change the (meta)data sharing practice of the microbial genomics community
- Enjoy the rest of the workshop!

Thank you for your attention

WWW.CIDGOH.CA

We are on a Coffee Break & Networking Session

Workshop Sponsors:

Canadian Centre for Computational Genomics

HPC4Health

OICR
Ontario Institute for Cancer Research

Ontario Genomics

GenomeCanada

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada