State of the Science: Next Generation Sequencing and Bioinformatics

Pushing Back the Frontiers of Science: A Review of This Year’s Literature
APHL 2023 ID Lab Con
Pathogen Genomics in Public Health

The NEW ENGLAND JOURNAL of MEDICINE

SPECIAL REPORT

Pathogen Genomics in Public Health

Gregory L. Armstrong, M.D., Duncan R. MacCannell, Ph.D., Jill Taylor, Ph.D., Heather A. Carleton, M.P.H., Ph.D., Elizabeth B. Neuhaus, Ph.D., Richard S. Bradbury, Ph.D., James E. Posey, Ph.D., and Marta Gwinn, M.D., M.P.H.

Pathogen Genomics in Public Health

Figure 1. Example of Sequencing for Outbreak Detection and Investigation.

An important purpose of surveillance for infectious diseases is to identify outbreaks for investigation and intervention. Discovering patterns in the epidemiologic data (i.e., finding common exposures among cases that cluster in time and location) can help distinguish outbreaks from the often much larger background of sporadic cases. Molecular subtyping has played an increasingly central role in this process through the detection of cases with isolates that share a common molecular “fingerprint.”

In this figure, we schematically represent surveillance data for a foodborne pathogen, *Salmonella enterica* serovar *Enteritidis*, reported from one region of the United States in 2018. In that year, some states in the region were already sequencing salmonella isolates in real time, and others had not yet started. In the three panels, each dot represents a case of *gastroenteritis* due to *S. enterica* serovar *Enteritidis*. Gray dots represent cases that were later determined to be “sporadic” (i.e., not linked to outbreaks), and colored dots represent cases that were eventually linked to outbreaks. The largest of these outbreaks (red dots) began as two distinct clusters of disease associated with restaurants in two different states. Whole-genome sequencing linked these two clusters together and to several other cases outside the region.

Panel A displays cases randomly, without regard to molecular subtyping. Panel B represents a grouping of cases based on results of pulsed-field gel electrophoresis, a molecular subtyping technology that U.S. public health agencies have used since the 1990s. In this example, pulsed-field gel electrophoresis was mostly successful at grouping cases from the largest (red) outbreak; however, the group includes many cases unrelated to the outbreak, which complicates the investigation and reduces the likelihood of finding the food source. Panel C shows that the finer resolution afforded by whole-genome sequencing was more effective in segregating the red outbreak cases from others. This gave investigators more confidence in the cluster definition and allowed them to focus on cases that were more likely part of the same outbreak. In this outbreak, epidemiologic investigation identified shell eggs as the likely source, which was quickly confirmed by isolating *S. enterica* serovar *Enteritidis* from the implicated eggs and showing that its whole-genome sequence matched that from the outbreak cases.

In addition to the outbreak in red, this panel shows four additional outbreaks. Cases in blue were part of a restaurant-associated outbreak linked to chicken in a single state. Two cases (purple) were linked to live poultry exposure as part of a much larger, multistate, multistrain outbreak that occurred mostly outside the region shown here. The 5 cases in light pink were investigated as an outbreak, but no food source was identified. The 15 cases in light orange occurred in a state where real-time whole-genome sequencing had not yet been implemented; their isolates were not sequenced until a later date, after the apparent outbreak had ended. This figure summarizes the relationships identified by whole-genome sequencing with the use of a simplified graph; in practice, however, the data would be represented as a phylogenetic tree, which contains additional detail that more precisely represents the relationships among sequences.
SARS-CoV-2

A new coronavirus associated with human respiratory disease in China

Fan Wu,1,2, Su Zhao,1,2, Bin Yu,1,2, Yan-Mei Chen,1,2, Wen Wang,1,2, Zhi-Gang Song,1,2, Yi Hu,1,2, Zhao-Wu Tao,1, Jun-Hua Tian,1, Yuan-Yuan Pei,1, Ming-Li Yuan,1, Yu-Ling Zhang,1, Fa-Hui Dai,1, Yi Liu,1, Qi-Min Wang,1, Jiao-Jiao Zheng,1, Lin Xu,1, Edward C. Holmes1,6 & Yong-Zhen Zhang1,6

Emerging infectious diseases, such as severe acute respiratory syndrome (SARS) and Zika virus disease, present a major threat to public health1–3. Despite intense research efforts, how, when and where new diseases appear are still a source of considerable uncertainty. A severe respiratory disease was recently reported in Wuhan, Hubei province, China. As of 23 January 2020, at least 1,975 cases had been reported since the first patient was hospitalized on 12 December 2019. Epidemiological investigations have suggested that the outbreak was associated with a seafood market in Wuhan. Here we study a single patient who was a worker at the market and who was admitted to the Central Hospital of Wuhan on 26 December 2019 while experiencing a severe respiratory syndrome that included fever, dizziness and a cough. Metagenomic RNA sequencing4 of a sample of bronchoalveolar lavage fluid from the patient identified a new RNA virus strain from the family Coronaviridae, which is designated here ‘WH-Human I’ coronavirus (and has also been referred to as ‘2019-nCoV’).

Phylogenetic analysis of the complete viral genome (29,903 nucleotides) revealed that the virus was most closely related (89.1% nucleotide similarity) to a group of SARS-like coronaviruses (genus BetaCoV, subgenus Sarbecovirus) that had previously been found in bats in China5. This outbreak highlights the ongoing ability of viral spill-over from animals to cause severe disease in humans.
SARS-CoV-2

https://artic.network/

About

Our goals

This project is developing an end-to-end system for processing samples from viral outbreaks to generate real-time epidemiological information that is interpretable and actionable by public health bodies.

Fast evolving RNA viruses (such as Ebola, SARS-CoV-2, influenza etc) continually accumulate changes in their genomes that can be used to reconstruct the epidemiological processes that drive the epidemic.

Based around a recently developed, single-molecule portable sequencing instrument, the Oxford Nanopore Technology MinION, we are creating a ‘lab-in-a-suitcase’ that can be deployed to remote and resource-limited locations. Targeting a wide-range of emerging viral diseases, the sequencing generation will be closely linked to the analysis platform to integrate these data and associated epidemiological knowledge to reveal the processes of transmission, virus evolution and epidemiological linkage with extremely rapid turn-around.

This real-time approach will provide actionable epidemiological insights within days of samples being taken from patients.
SARS-CoV-2 Sequencing on Illumina MiSeq Using ARTIC Protocol: Part 1 - Tiling PCR V.1

In 2 collections

Joel Sevinsky1, Arian Nassiri2, Erin Young3, Heather Blankenship4, Kevin Libuit2, Kelly Oakeson3, Lauren Turner2, StaPH-B Consortium5

1Theagien Genomics; 2Virginia Division of Consolidated Laboratory Services; 3Utah Public Health Laboratory; 4Michigan Department of Health and Human Services; 5State Public Health Bioinformaticians

Benchmark datasets for SARS-CoV-2 surveillance bioinformatics

Lingli Xiaoli1, Jill V. Hagey1, Daniel J. Park1, Christopher A. Gulvik1, Erin L. Young1, Nabil-Fared Alkilani1, Adrian Lawsin1, Norman Hassell1, Kristen Knipe1, Kelly F. Oakeson2, Adam C. Retchless, Mignon Shakes1, Chien-Chi Lee1, Patrick Chiang1, Andrew J. Page1, Benjamin J. Metcalf1, Michelle Su1, Jessica Rowell1, Sathaw Vidyasekar1, Clinton R. Paden1, Andrew D. Huang1, Dawn Roelli1, Ketan Patel1, Kathryn Winglee1, Michael R. Weigand1 and Lee S. Katz1

1Strain Surveillance and Emerging Variant Team, Centers for Disease Control and Prevention, Atlanta, GA, United States of America
2Broad Institute of MIT and Harvard, Cambridge, MA, United States of America
3Utah Public Health Laboratory, Salt Lake City, UT, United States of America
4Quadram Institute Bioscience, Norwich Research Park, Norwich, United Kingdom
5Resilience Division, Los Alamos National Laboratory, Los Alamos, NM, United States of America
6SARS-CoV-2 Emerging Variant Sequencing Project Dry Lab Group Laboratory and Testing Task Force

Low SARS-CoV-2 Transmission in Elementary Schools — Salt Lake County, Utah, December 3, 2020–January 31, 2021

Rebecca B. Hershow, PhD1,2,3; Karen Wu, DVM1,2,4; Nathaniel M. Lewis, PhD1,2,3; Alison T. Milne, EdD1; Dustin Currie, PhD2,3; Amanda R. Smith, PhD1,2,3; Spencer Liew, MD4; Brian Oslanski, MD3; Erin L. Young, PhD2; Brandi Freeman, PhD2; Noah Schwartz, MD2,3; Bobbi Beyrer, MPH3; Catherine Espinosa, MPH1; Yvonne Nakawumu, PhD1; Elizabeth Gara, MPH1; Oliva Almeda, MPH1; Winston E. Abra, MD3; Daniel C. Ehlan, ScD1; Keith Waines, MD3; Mary Hill, MPHT3; lone Rieh3; MKH4; Kyle Osborn, RN; Rebecca L. Kirking, MD3; Angela Dunn, MD3; Snigdha Vallabhshani, MD4; Adam L. Hers4, MD3; Victoria T. Chu1, MD2,3

Hershow, R et al. Low SARS-CoV-2 Transmission in Elementary Schools — Salt Lake County, Utah, December 3–January 31, 2021. MMWR. https://doi.org/10.15585/mmwr.mm7012e3

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

Xianding Deng1,2, Wei Gu1,2, Scott Federman2, Louis du Plessis4, Oliver G. Pybus1,4, Nuno R. Faria1,3, Candace Wang1,2, Guixia Yu1,2, Brian Bushnell1, Chao-Yang Pan1, Hugo Guevara1, Alicia Sotomayor-Gonzalez1,2, Kelsey Zorn1, Allam Gheez2, Venice Serveltta2, Elaine Hsu2, Steve Miller1, Trevor Bedford2, Alexander L. Greninger1, Pavitra Roychoudhury3, Lea M. Starks1,2, Michael Fumadera1, Helen Y. Chu1,2, Jay Shendure1,3, Keith K. Jerome1,3, Catie Anderson1, Karthik Ganagavarma1, Mark Zeilis1, Emily Spencer1, Kristian G. Andersson1, Duncan MacCannell1, Clinton R. Paden1, Yan Li1, Jing Zhang1, Suxiang Tong1, Gregory Armstrong1, Scott Morrow1, Matthew Willins1, Bela T. Matyas1, Sundari Mase1, Olivia Kasire1, Maggie Park1, Godfried Maxinde1, Curtis Chan1, Alexander T. Yu1, Shua J. Chai1, Elsa Villarino1, Brandon Bonin1, Debora A. Wadford1, Charles Y. Chu1,2,4,6

The coronavirus disease 2019 (COVID-19) pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has spread globally, with >365,000 cases in California by 17 July 2020. We investigated the genomic epidemiology of SARS-CoV-2 in Northern California from late January to mid-March 2020, using samples from 36 patients spanning nine counties and the Grand Princess cruise ship. Phylogenetic analyses revealed the cryptic introduction of at least seven different SARS-CoV-2 lineages into California, including epidemic WAI strains associated with Washington state, with lack of a predominant lineage and limited transmission among communities. Lineages associated with outbreak clusters in two counties were defined by a single base substitution in the viral genome. These findings support contact tracing, social distancing, and travel restrictions to contain the spread of SARS-CoV-2 in California and other states.
Open Pathogen Genomics

Ten recommendations for supporting open pathogen genomic analysis in public health

Allison Black1,2, Duncan R. MacCannell3, Thomas R. Sibley2 and Trevor Bedford1,2

Increasingly, public-health agencies are using pathogen genomic sequence data to support surveillance and epidemiological investigations. As access to whole-genome sequencing has grown, greater amounts of molecular data have helped improve the ability to detect and track outbreaks of diseases such as COVID-19, investigate transmission chains and explore large-scale population dynamics, such as the spread of antibiotic resistance. However, the wide adoption of whole-genome sequencing also poses new challenges for public-health agencies that must adapt to support a new set of expertise, which means that the capacity to perform genomic data assembly and analysis has not expanded as widely as the adoption of sequencing itself. In this Perspective, we make recommendations for developing an accessible, unified informatic ecosystem to support pathogen genomic analysis in public-health agencies across income settings. We hope that the creation of this ecosystem will allow agencies to effectively and efficiently share data, workflows and analyses and thereby increase the reproducibility, accessibility and auditability of pathogen genomic analysis while also supporting agency autonomy.

Open Pathogen Genomics

1. Support data hygiene and interoperability by developing and adopting a consistent data model.
2. Strengthen application programming interfaces.
3. Develop guidelines for management and stewardship of genomic data.
4. Make bioinformatics pipelines fully open-source and broadly accessible.
5. Develop modular pipelines for data visualization and exploration.
6. Improve the reproducibility of bioinformatics analyses.
7. Utilize cloud computing to improve the scalability and accessibility of bioinformatics analyses.
8. Support new infrastructure and software development demands with an expanded technical workforce.
9. Improve the integration of genomic epidemiology with traditional epidemiology.
10. Develop best practices to support open data sharing.
Open Pathogen Genomics

Pipelines for data processing

- FASTQ files transferred to a raw data repository
  - SRA
  - BaseSpace
  - Pathogenwatch

- Assembly pipeline A
- Assembly pipeline B
- Assembly pipeline C

Multiple possible assembly pipelines
Vary depending on pathogen or task

Local public or private databases holding sample metadata and accession ID for assembled genome in assembly repository

- Metadata DB

Private database on secure infrastructure housing patient-identifiable metadata

- Identifiable information DB

Pipelines for data analysis and visualization

- Assembled genomes transferred to an assembly repository
  - NCBI
  - ENA
  - Pathogenwatch

API

API sources genomic assembly and necessary metadata for visualization or analysis

- Genome
  - and metadata

- BWA
  - RAxML

- Upload to tree DB

- Nextstrain API
  - auspice
  - JSON

- Zika Nextstrain

- TB AMR API

- Predict TB AMR

- PDF report

- Genome
  - and secure EpiData

- Secure HIV trace
Machine Learning & AI

REVIEW ARTICLE

Machine learning and applications in microbiology

Stephen J. Goodswen¹,†, Joel L. N. Barratt², Paul J. Kennedy³,* , Alexa Kaufer¹, Larissa Calarco¹ and John T. Ellis¹,*

¹School of Life Sciences, University of Technology Sydney (UTS), 15 Broadway, Ultimo, NSW 2007, Australia, ²Parasitic Diseases Branch, Division of Parasitic Diseases and Malaria, Center for Global Health, Centers for Disease Control and Prevention, 1600 Clifton Rd, Atlanta, GA 30333, USA and ³School of Computer Science, Faculty of Engineering and Information Technology and the Australian Artificial Intelligence Institute, University of Technology Sydney (UTS), 15 Broadway, Ultimo, NSW 2007, Australia


Machine Learning Advances in Microbiology: A Review of Methods and Applications

Yiru Jiang, Jing Luo, Danqing Huang, Ya Liu* and Dan-dan Li*

Machine Learning & AI

Figure 1. A hierarchical perspective of machine learning within the artificial intelligence framework. Machine Learning (ML)—an application that provides the capacity to automatically learn and improve from experience; Data mining—aims to find useful information from large volumes of data using computer algorithms (not covered in review); Supervised—ML using labelled training data (it can be compared to a human learning in the presence of a supervisor); Unsupervised—ML with unlabelled data (comparable to human learning without a supervisor); Reinforcement learning—focuses on taking a set of actions to maximise reward given a particular environment (requires no training data) (not covered in review); Classification—the task of predicting a discrete class label. A classification model is often referred to as a classifier. Examples of binary classification labels: yes/no, 1/0, vaccine/non-vaccine; and examples of multi-class labels: A, B, C, D for student grades; and teacher, student, secretary, and principal for a school classification; Regression—the task of predicting a continuous quantity; Clustering—an exploratory or descriptive approach in contrast to a predictive approach as in classification and regression; Association rule mining—identifies patterns of association between different variables e.g. movie suggestion, market basket analysis.
Applications of Machine Learning to the Problem of Antimicrobial Resistance: an Emerging Model for Translational Research

Machine Learning & AI

Machine learning is being used in various ways to predict antibiotic resistance, which is a major concern in public health. Here are some examples:

1. Genomic analysis: Machine learning algorithms can be trained to analyze the genomic data of bacterial strains and predict their antibiotic resistance patterns. This approach can identify specific genetic mutations or features that contribute to antibiotic resistance.
2. Clinical data analysis: Machine learning algorithms can be trained to analyze clinical data, such as patient demographics, medical histories, and lab results, to predict the likelihood of antibiotic resistance. This approach can help clinicians make more informed decisions about which antibiotics to prescribe.
3. Predictive modeling: Machine learning algorithms can be used to build predictive models that forecast the spread of antibiotic-resistant bacteria in a population. This approach can help public health officials allocate resources and implement targeted interventions to prevent outbreaks.
4. Drug discovery: Machine learning algorithms can be used to design new antibiotics that are less likely to induce resistance. This approach can help overcome the challenges of traditional drug discovery methods and accelerate the development of effective treatments.

Overall, machine learning is a promising tool in the fight against antibiotic resistance, and its applications are likely to expand as more data becomes available and new algorithms are developed.

Machine learning is increasingly used in epidemiology to analyze complex datasets and identify patterns and insights that can inform public health policy and interventions. Here are some ways in which machine learning is being used in epidemiology:

1. Disease surveillance: Machine learning algorithms can be used to analyze large datasets of clinical and demographic data to identify outbreaks of infectious diseases in real-time. This approach can help public health officials detect and respond to outbreaks more quickly.
2. Disease prediction: Machine learning algorithms can be used to build predictive models that forecast the likelihood of disease outbreaks in a population. This approach can help public health officials allocate resources and implement targeted interventions to prevent outbreaks.
3. Risk factor analysis: Machine learning algorithms can be used to analyze data on various risk factors, such as environmental exposure, genetic predisposition, and lifestyle factors, to identify the factors that contribute to disease risk.
4. Drug discovery: Machine learning algorithms can be used to design new drugs and optimize drug therapies based on patient characteristics and disease progression. This approach can help improve treatment outcomes and reduce the risk of adverse drug reactions.
5. Health behavior analysis: Machine learning algorithms can be used to analyze data on health behaviors, such as diet and exercise, to identify patterns and insights that can inform public health campaigns and interventions.

Overall, machine learning is a valuable tool in epidemiology, and its applications are likely to expand as more data becomes available and new algorithms are developed.
SEQUENCE ALL THE THINGS

Questions?