Published in final edited form as:
N Engl J Med. 2019 December 26; 381(26): 2569–2580. doi:10.1056/NEJMsr1813907.

Pathogen Genomics in Public Health

Gregory L. Armstrong, M.D.,
National Center for Emerging and Zoonotic Infectious Diseases

Duncan R. MacCannell, Ph.D.,
National Center for Emerging and Zoonotic Infectious Diseases

Heather A. Carleton, M.P.H., Ph.D.,
National Center for Emerging and Zoonotic Infectious Diseases

Elizabeth B. Neuhaus, Ph.D.,
National Center for Immunization and Respiratory Diseases

Richard S. Bradbury, Ph.D.,
Center for Global Health (R.S.B.)

James E. Posey, Ph.D.,
National Center for HIV/AIDS, Viral Hepatitis, STD and TB Prevention

Jill Taylor, Ph.D.,
Centers for Disease Control and Prevention, Wadsworth Center, New York State Department of Health

Marta Gwinn, M.D., M.P.H.
CFOL International

Abstract

Rapid advances in DNA sequencing technology (“next-generation sequencing”) have inspired optimism about the future potential of human genomics for “precision medicine.” Meanwhile, pathogen genomics is already delivering “precision public health” via more effective foodborne illness outbreak investigations, better targeted tuberculosis control, and more timely and granular influenza surveillance to inform vaccine strain selection. In this article, we describe how public health agencies are rapidly adopting pathogen genomics to improve their effectiveness in almost all domains of infectious disease. This momentum is likely to continue, given ongoing development in sequencing and sequencing-related technologies.

An important transformation is underway in public health. Next-generation sequencing (NGS, also called “high-throughput sequencing”) is reshaping communicable disease surveillance, allowing for earlier detection and more precise investigation of outbreaks. NGS helps characterize microbes more effectively and offers new insights into their ecology and

Corresponding author: Gregory L. Armstrong, MD, Centers for Disease Control and Prevention, 1600 Clifton Rd., NE, Atlanta, GA 30329, 404.639.0422, garmstrong@cdc.gov.

transmission. The plethora of sequence data provides raw material for research and development of new diagnostics and therapeutics.

This article describes how pathogen genomics has been changing public health in the United States and globally.

Adapting NGS to Public Health Use

The NGS era began with the commercial release of massively parallel pyrosequencing in 2005, the first fundamental advance in sequencing technology since the invention of Sanger sequencing in the 1970s. [1,2] In the early years, NGS efficiency improved rapidly, with sequencing costs falling by as much as 80% year-over-year. [1,3] In public health, these developments were both exciting—because of the myriad potential applications, including bacterial whole-genome sequencing (WGS)[4]—and intimidating—because of the barriers: implementing NGS would require investment in sequencing equipment as well as high-performance computing infrastructure to move, store and analyze large volumes of sequence data. Equally important was the need to integrate bioinformatics, a discipline new to public health.

Public Health England was an early leader in the use of NGS at a national scale, particularly for tuberculosis[5,6] and bacterial foodborne disease surveillance.[7,8] In the United States, CDC was a late adopter,[9] but is now applying the technology broadly, due largely to the Advanced Molecular Detection (AMD) program, a $30 million dollar per year initiative established by Congress in 2013 to bring NGS and other innovative laboratory technologies to bear against infectious disease threats, first at CDC and then in state and local public health departments nationwide.

Applications of Pathogen Genomics

Today, pathogen genomics is part of almost every infectious disease program at CDC.[10] Some applications of NGS that serve specialized purposes, such as reference testing, are in use only at CDC, while others drive entire domestic surveillance systems. Below, we provide examples to highlight the value of NGS technology for public health (Box).

Bacterial foodborne illness.

In the mid-1990s, U.S. foodborne disease programs first began applying standardized molecular subtyping—pulsed-field gel electrophoresis (PFGE)—to bacterial pathogens as part of routine surveillance, leading to a fundamental change in how outbreaks were identified and investigated. The resulting national network, “PulseNet”, now includes more than 80 public health laboratories.[11] Before PulseNet, outbreaks were difficult to detect and solve unless they were large or very geographically and temporally focused. For example, during the 20-year period before PulseNet, only five outbreaks of listeriosis (0.25 per year) were solved (i.e., with a food source identified),[12] with a mean of 54 cases per outbreak. In the 5-year period after PulseNet began, 11 outbreaks were identified (2.2 per year) with a median of 5 cases per outbreak.[12] Routine use of PFGE had a similar impact on the detection and response to other foodborne bacteria, particularly *Salmonella*[13] and Shiga toxin-producing *Escherichia coli* (STEC).[14–16]

PulseNet has now transitioned from PFGE to WGS.[17,18] Partners in this effort include the U.S. Food and Drug Administration, whose GenomeTrakr system[19,20] performs WGS of food and environmental isolates, the U.S. Department of Agriculture, and the National Center for Biotechnology Information (NCBI).

Compared with PFGE, WGS offers a vastly finer resolution: typically, a three- to six-million base-pair sequence, in contrast to a gel pattern with ten to twenty bands that reflect changes in small parts of the genome. WGS data are inherently digital, standardized, and much less dependent on the choice of laboratory protocol. The results reveal evolutionary relationships between bacterial isolates, allowing a better understanding of transmission and links between cases (Figure 1). WGS can also predict phenotypic characteristics, such as virulence, serotype and antimicrobial resistance.[15,21–26] Costs for WGS (around $200 to $250 per isolate, including consumables, labor, equipment, maintenance and overheads) are currently still higher than those for PFGE (around $100 according to the same analysis), although the higher costs may be partly or entirely offset by eliminating the need for traditional phenotyping assays. In addition, advances in sequencing technology and laboratory automation may further reduce WGS costs.

It is too early to know how the transition to WGS will affect U.S. surveillance for more common foodborne pathogens, such as *Salmonella* and STEC; however, early experience with *Listeria* surveillance, which switched to routine WGS in 2014, has been encouraging. In the first three years of WGS (September 2013 through August 2016), 18 outbreaks of listeriosis were solved (6 per year) with a median of 4 cases per outbreak.[27] In the United Kingdom, where WGS has been in routine use for STEC since at least 2015, the number of clusters detected has doubled.[28]

Tuberculosis.

Since the 1990s, several DNA fingerprinting technologies have proven useful for subtyping *Mycobacterium tuberculosis* (MTB).[29] Identifying closely-related strains allows health department investigators to detect clusters of cases that may be linked to recent transmission —cases requiring more intense investigation and possible intervention.[30] WGS offers much finer resolution subtyping than older technologies and thus more confidence in the inferred relationships among cases. After using WGS selectively for several years, the U.S. tuberculosis control program has now scaled up to sequence isolates from all culture-confirmed cases nationwide. In California, WGS allowed public health workers to refute more than half of suspected outbreaks initially identified by conventional genotyping, saving time and resources (personal communication, Tambi Shaw, California Department of Public Health). Early experience in U.K.,[5] Canadian,[31,32] and Dutch[33] tuberculosis programs has also confirmed that WGS supports more effective investigations by more accurately defining outbreaks[5,31,33], providing insights into transmission dynamics,[34] and sometimes suggesting the presence of previously unidentified cases or possible "super-spreaders" that should be prioritized for isolation and treatment.[5,31] WGS may also indicate whether recurrent cases are due to reactivation or reinfection, information useful in evaluating program effectiveness.[35]

The ability to prioritize case-investigations may also be of use in the high-incidence, low- and middle-income countries which carry the world's heaviest burden of tuberculosis.[36] In these countries, however, a different application of NGS—sequencing of MTB directly from sputum—could have an even more important role.[37] Direct sequencing of MTB from smear-positive sputum samples is already feasible in research settings[38–40] but is too expensive and cumbersome for routine clinical and public health use. If it can be made practical and cost-effective, this approach will enable rapid inference of drug susceptibility, which is already quite accurate for most first-line drugs, and will improve as more data become available.[37,41] In addition to supporting prompt treatment with appropriate drugs, NGS will reduce the need for routine phenotypic testing, which is complex, slow, and difficult to maintain in resource-limited laboratory settings.

In the meantime, an intermediate strategy is already practical in high-income countries: WGS directly from early positive cultures, providing information on drug susceptibility weeks before the results of traditional tests are available.[38] Laboratories in both the New York State Department of Health and Public Health England[37] have received regulatory approval to forego traditional drug-susceptibility testing of isolates predicted by WGS to be susceptible to all four first-line drugs—approximately 70–80% of all isolates.[37]

Another promising strategy is targeted amplicon sequencing of selected mycobacterial genes or marker sequences.[42] To remain relevant over time, any sequence-based method for inferring drug susceptibility must rely on the continuous updating of databases with correlated genotypic and phenotypic data.[37]

Influenza.

The selection of seasonal influenza vaccine candidate strains is a complex, global undertaking, involving massive surveillance efforts from dozens of countries and contributing organizations. The World Health Organization (WHO) convenes international experts twice yearly to review information on circulating influenza strains and, based on that information, to recommend components of the Northern and Southern Hemisphere influenza vaccines.[43] CDC contributes to this process by overseeing characterization of 4,000 to 10,000 influenza specimens each year.[44]

The traditional method for characterizing influenza strains begins with viral culture, which is increasingly challenging for certain strains, particularly H3N2.[45] Two or more passages through culture are often required, during which some adaptation of the virus may occur. Next, a few viral isolates are selected for phenotyping, which previously included antigenic characterization and consensus (Sanger) sequencing of selected genes. These steps are time-consuming and labor-intensive.

NGS now enables a more efficient "sequence-first" approach, in which original specimens are subjected directly to whole-genome reverse-transcriptase PCR, followed by sequencing.[45–47] These sequence data provide a highly granular view of viral emergence and allow for a more parsimonious selection of viruses for phenotypic characterization, including antigenic analysis and susceptibility to antiviral agents. This approach is not only faster but also more informative. For example, detailed phylogenetic information on all viral segments provides a richer picture of how influenza viruses are diversifying to evade existing immunity, and deep sequencing (sequencing many copies of the genome from the same sample) can reveal the presence of drug-resistant minor variants not reflected in the consensus sequence. Sequencing cannot completely replace traditional phenotyping, but sequencing first, using clinical specimens, allows it to be done more selectively.

Influenza NGS data are now routinely reviewed at the biannual WHO consultations and have already affected vaccine decision-making in at least two major instances, most recently contributing to a change in the A(H3N2) vaccine component to target a newly emerging clade.[45,48,49] NGS data are also used for forecasting the relative importance of emerging strains and risk assessment,[50–52] characterizing viruses used in vaccine effectiveness studies,[53] and informing treatment for patients infected with high pandemic risk viruses such as H7N9.[54,55]

Parasitic diseases.

Diagnosis of many parasitic diseases continues to rely on microscopy, a nineteenth-century technology that is operator-dependent and resistant to automation. PCR and other diagnostic techniques (e.g., serology) have been developed for many common parasites but require separate tests for each suspected pathogen.

CDC's parasitic diseases laboratories are developing a new type of diagnostic test based on the targeted and direct sequencing of eukaryotic housekeeping genes. This approach should enable the accurate detection of all known potential parasitic agents present in a blood sample with a single test. Early validation data suggest that this novel assay is at least as sensitive as standard PCR for parasites found in blood.[56] Further validation and development is ongoing, with plans to add more targets and to adapt the assay to more complex samples, including tissue and stool.

PCR amplification and NGS of specific genes is an effective means of identifying drug-resistance in malaria parasites.[57] Because such testing takes 2–3 days and is still somewhat expensive, its use in routine patient care remains limited. For surveillance at the country or regional level, however, it is a cost-effective and efficient way to assess drug resistance and to target treatment recommendations more precisely. NGS-based protocols may also be useful for assessing the intensity of transmission by gauging multiplicity of infection among residents in malaria-endemic areas.[58]

Cyclospora cayetanensis causes outbreaks of foodborne diarrheal disease in the US every year. The organism's limited genotypic variability and its inability to be propagated in the laboratory have confounded the development of effective genotyping methods for surveillance. CDC has developed a method to extract *C. cayetanensis* directly from stool and used it to produce whole genomes of multiple isolates.[59] Several promising genotyping targets have been identified and their translation into a functional and discriminatory genotyping tool is showing promise.[60] The genotyping tool and associated analysis system underwent further testing and validation during 2018's unusually active summer cyclosporiasis season.

Other applications.

NGS is applicable across the spectrum of important pathogens in public health. For Legionnaires' disease, for example, finer subtyping has been useful for investigating and responding to outbreaks.[61,62] Eventually, insights from NGS into *Legionella* ecology and persistence in water systems could improve prevention.[63] For hospital acquired infections, NGS is proving to be an invaluable tool for identifying and investigating outbreaks[64,65] and also for better understanding transmission both at the hospital[65] and community[64,66] level. For HIV, genetic sequence data generated for clinical purposes can be analyzed to identify potential clusters for early public-health intervention;[67,68] user-friendly tools[69] now allow state and local health departments to make use of these data. Community-level molecular surveillance for hepatitis C clusters has also proven useful.[70]

Other applications for NGS in public health include tracking the emergence of antimicrobial resistance and novel resistant pathogens such as *Candida auris*,[66,71] tracking insecticide resistance in mosquito vectors of disease,[72,73] monitoring streptococcal pathogens,[74,75] investigating potential clusters of meningitis,[76] and many more.

Sequencing to support outbreak response.

An important subset of NGS applications in public health involves outbreak response.[77] For bacterial foodborne disease, for example, sequencing is central to detecting outbreaks, investigating cases, confirming the implicated food and tracing it back to its source. A very different example comes from the latter phases of the 2014-'16 Ebola outbreak in Guinea, in which sequencing was useful in ascertaining the likely source of infection in "outlier" cases —those with no known connection to other cases. Each outlier raised troubling questions: Did the case represent another introduction from an animal reservoir? Was it attributable to sexual transmission? Had a long chain of transmission been missed, suggesting serious gaps in surveillance? Each of these would require a different response. Fortunately, in this outbreak, sequences from outliers were consistently closely related to those in the known outbreak zone, and the response team was able to remain focused on stopping transmission there.[10,78]

In contrast, sequencing Zika virus during its emergence in the Americas was not useful for responding to individual cases. However, sequence data were critical for developing diagnostics and vaccines, to better understand the evolution of the epidemic,[79] and recently to support evidence of an undetected outbreak in Cuba.[80] Analysis of Ebola virus genomes after the West Africa outbreak also provided useful insights into the virus's spread.[81]

Where to Go Next?

Six years into the AMD program, NGS is now central to US public health programs for monitoring, controlling, and preventing infectious diseases. Progress is currently needed in several areas:

Metagenomics.

Sequencing, particularly bacterial WGS, often requires pure cultures of the organisms, which are increasingly difficult to obtain as clinical laboratories expand the use of highly multiplexed, syndrome-based, culture-independent diagnostic tests.[82,83] Most immediately, this trend is adversely affecting enteric disease surveillance as clinical laboratories move away from culture.[84] One solution may be to bypass culture, sequencing pathogen genomes directly from specimens.[38,85] Although already feasible for selected organisms in particular specimens, the methods are not yet practical for routine use.

Data integration and data science.

In public health settings, laboratory and epidemiologic data are often stored and managed separately. Until recently, these data could be brought together for analysis without loss of information: laboratory data such as a positive/negative result, a titer, or even a PFGE pattern identifier could be imported into an epidemiologic database and analyzed with traditional statistical tools. With pathogen genomic data, this is no longer true: these data need to be integrated to realize the full value of both.[31,86,87]

Fortunately, academic research is addressing this challenge,[88] producing tools for visualizing and analyzing epidemiologic and phylogenetic data together, such as Microreact (microreact.org),[89] Nextstrain (nextstrain.org)[90] or the Interactive Tree of Life (itol.embl.de).[91] More broadly, the emerging field of data science offers novel approaches for integrating, analyzing and visualizing increasingly diverse public health data.[20,92]

Software to facilitate NGS workflows.

Complex workflows are required to manage the sequencing process, analyze raw sequence data, store processed data and integrate it with epidemiologic data, and finally, share information securely. In a single academic lab, it's feasible for a bioinformatician to accomplish this, but in a network of public health laboratories, it can be more practical to manage much of this centrally. Bioinformatics tools are available to accomplish the basic steps of assembling raw sequence data into genomes and "pipelines" using these tools are available to automate core processes, such as validating quality, assembling a genome, and inferring phenotypes. However, user-friendly tools for managing workflows and integrating data are often lacking. In other instances, too many tools exist and it is not clear which ones will survive the test of time.

Challenges

Workforce transformation, both among microbiologists and epidemiologists, while perhaps the most obvious hurdle, is in some ways the least challenging due to enthusiasm about genomics. Even recruiting and retaining bioinformaticians has not been as difficult as initially anticipated. Among the first 27 recruits into the CDC/Association of Public Health Laboratories Bioinformatics Fellowship, 19 (70.4%) are still working in public health, despite a competitive marketplace. Those who stay in public health often cite the opportunity to have positive social impact as a key motivator.

Costs represent a more difficult issue. For now, sequencing is often more expensive than traditional subtyping. In addition, to minimize per-sample costs, sequencing technologies such as the MiSeq platform (Illumina, San Diego, Calif.) may require batch sizes of 15 samples or more, which can extend turn-around times beyond the 36 hours needed for the sequencing itself. Clearly, during a fast-moving outbreak, such delays are undesirable. Single-molecule, long-read sequencing technology, most notably the MinION (Oxford Nanopore Technologies, Oxford, UK), already has niche uses within public health because of its portability, but also has the potential to reduce batch-size needs and turn-around times.[78]

Data Openness

In both academia and public health, pathogen genomics is ushering in a new era in data openness. In the United States, local, state and federal agencies are uploading data on bacterial foodborne pathogens,[20] influenza[45] and other pathogens to public databases hosted by NCBI (ncbi.nlm.nih.gov/pathogens), making these data available in near real time. These groups also contribute to other global databases, such as the Relational TB Sequencing Data Platform ("ReSeqTB", reseqtb.org) and the Global Initiative on Sharing All Influenza Data (GISAID, gisaid.org), established to promote international exchange of sequence and other data. In addition to enabling secondary uses of data, openness also encourages collaboration among public health organizations, academia, and industry. Nevertheless, openness can never be complete and unconditional: public health agencies have always been vigilant guardians of confidentiality, and pathogen genomic data are released only after careful consideration of risks.[87,93,94]

Conclusion

NGS and bioinformatics are transforming the response to infectious disease outbreaks, providing new insights into disease emergence and transmission, expediting pathogen characterization, and promoting data sharing (Figure 2). In public health, sequencing is already routine in many core domains, including foodborne bacterial pathogens, tuberculosis, influenza and antimicrobial resistance. These developments are taking place within a rapidly evolving technology landscape: NGS is becoming more automated, efficient, and accurate and related technologies, such as systems for highly multiplexed DNA amplification, are advancing.

Public health workforce development is central to this process. Microbiologists need a strong knowledge base in microbial genomics. Epidemiologists need the skills and tools to translate genomic data into public-health action. Both groups need to grasp the basic vocabulary of bioinformatics. Public health should strive to attract professionals with broadly applicable data-science skills. For anyone considering a career in public health, this is an exciting time to jump in.

Acknowledgments

Disclaimer

The findings and conclusions in this report are those of the authors and do not necessarily represent the views of the CDC/the Agency for Toxic Substances and Disease Registry. Use of trade names and commercial sources is for identification only and does not imply endorsement by the Office of Advanced Molecular Detection, National Center for Emerging and Zoonotic Diseases, CDC, the Public Health Service, or the US Department of Health and Human Services.

References

1. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 2016;17:333–51. [PubMed: 27184599]
2. MacCannell D. Platforms and analytical tools used in nucleic acid sequence-based microbial genotyping procedures. Microbiol Spectr 2019;7:AME-0005–2018.
3. DNA Sequencing Costs: Data. National Human Genomics Research Institute, 2018 (Accessed 08/19/2018, 2018, at https://www.genome.gov/27541954/dna-sequencing-costs-data/.)
4. Koser CU, Ellington MJ, Cartwright EJ, et al. Routine use of microbial whole genome sequencing in diagnostic and public health microbiology. PLoS Pathog 2012;8:e1002824.
5. Walker TM, Ip CL, Harrell RH, et al. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis 2013;13:137–46. [PubMed: 23158499]
6. Satta G, Lipman M, Smith GP, Arnold C, Kon OM, McHugh TD. Mycobacterium tuberculosis and whole-genome sequencing: how close are we to unleashing its full potential? Clin Microbiol Infect 2018;24:604–9. [PubMed: 29108952]
7. Mook P, Gardiner D, Verlander NQ, et al. Operational burden of implementing Salmonella Enteritidis and Typhimurium cluster detection using whole genome sequencing surveillance data in England: a retrospective assessment. Epidemiol Infect 2018;146:1452–60. [PubMed: 29961436]
8. Jenkins C, Dallman TJ, Grant KA. Impact of whole genome sequencing on the investigation of food-borne outbreaks of Shiga toxin-producing Escherichia coli serogroup O157:H7, England, 2013 to 2017. Euro Surveill 2019;24.
9. Blue Ribbon Panel--Future Strategies for Bioinformatics in CDC's Infectious Disease Laboratories. Centers for Disease Control and Prevention, 2011 at https://www.cdc.gov/amd/pdf/bioinformatics-panel-report.pdf.)
10. Gwinn M, MacCannell D, Armstrong GL. Next-generation sequencing of infectious pathogens. JAMA 2019;321:893–4. [PubMed: 30763433]
11. Ribot EM, Hise KB. Future challenges for tracking foodborne diseases: PulseNet, a 20-year-old US surveillance system for foodborne diseases, is expanding both globally and technologically. EMBO Rep 2016;17:1499–505. [PubMed: 27644260]
12. Cartwright EJ, Jackson KA, Johnson SD, Graves LM, Silk BJ, Mahon BE. Listeriosis outbreaks and associated food vehicles, United States, 1998–2008. Emerg Infect Dis 2013;19:1–9. [PubMed: 23260661]
13. Crowe SJ, Green A, Hernandez K, et al. Utility of combining whole genome sequencing with traditional investigational methods to solve foodborne outbreaks of Salmonella infections associated with chicken: a new tool for tackling this challenging food vehicle. J Food Prot 2017;80:654–60. [PubMed: 28294686]
14. Berenger BM, Berry C, Peterson T, et al. The utility of multiple molecular methods including whole genome sequencing as tools to differentiate Escherichia coli O157:H7 outbreaks. Euro Surveill 2015;20:1–11. [PubMed: 26132766]
15. Chattaway MA, Dallman TJ, Gentle A, et al. Whole genome sequencing for public health surveillance of Shiga toxin-producing Escherichia coli other than serogroup O157. Front Microbiol 2016;7:1–4. [PubMed: 26834723]
16. Joensen KG, Scheutz F, Lund O, et al. Real-time whole-genome sequencing for routine typing, surveillance, and outbreak detection of verotoxigenic Escherichia coli. J Clin Microbiol 2014;52:1501–10. [PubMed: 24574290]
17. Carleton HA. Whole-genome sequencing is taking over foodborne disease surveillance. Microbe 2016;11:311–7.

18. Lindsey RL, Pouseele H, Chen JC, Strockbine NA, Carleton HA. Implementation of whole genome sequencing (WGS) for identification and characterization of Shiga toxin-producing Escherichia coli (STEC) in the United States. Front Microbiol 2016;7:766. [PubMed: 27242777]

19. Allard MW, Strain E, Melka D, et al. Practical value of food pathogen traceability through building a whole-genome sequencing network and database. J Clin Microbiol 2016;54:1975–83. [PubMed: 27008877]

20. Allard MW, Bell R, Ferreira CM, et al. Genomics of foodborne pathogens for microbial food safety. Curr Opin Biotechnol 2018;49:224–9. [PubMed: 29169072]

21. McDermott PF, Tyson GH, Kabera C, et al. Whole-genome sequencing for detecting antimicrobial resistance in nontyphoidal Salmonella. Antimicrob Agents Chemother 2016;60:5515–20. [PubMed: 27381390]

22. Tyson GH, Zhao S, Li C, et al. Establishing genotypic cutoff values to measure antimicrobial resistance in Salmonella. Antimicrob Agents Chemother 2017;61:e02140–16.

23. Tyson GH, McDermott PF, Li C, et al. WGS accurately predicts antimicrobial resistance in Escherichia coli. J Antimicrob Chemother 2015;70:2763–9. [PubMed: 26142410]

24. Joensen KG, Tetzschner AM, Iguchi A, Aarestrup FM, Scheutz F. Rapid and easy in silico serotyping of Escherichia coli isolates by use of whole-genome sequencing data. J Clin Microbiol 2015;53:2410–26. [PubMed: 25972421]

25. Bale J, Meunier D, Weill FX, dePinna E, Peters T, Nair S. Characterization of new Salmonella serovars by whole-genome sequencing and traditional typing techniques. J Med Microbiol 2016;65:1074–8. [PubMed: 27481354]

26. Ashton PM, Nair S, Peters TM, et al. Identification of Salmonella for public health surveillance using whole genome sequencing. PeerJ 2016;4:e1752.

27. Besser J, Carleton HA, Gerner-Smidt P, Lindsey RL, Trees E. Next-generation sequencing technologies and their application to the study and control of bacterial infections. Clin Microbiol Infect 2018;24:335–41. [PubMed: 29074157]

28. Dallman TJ, Byrne L, Ashton PM, et al. Whole-genome sequencing for national surveillance of Shiga toxin-producing Escherichia coli O157. Clin Infect Dis 2015;61:305–12. [PubMed: 25888672]

29. Guthrie JL, Gardy JL. A brief primer on genomic epidemiology: lessons learned from Mycobacterium tuberculosis. Ann N Y Acad Sci 2017;1388:59–77. [PubMed: 28009051]

30. Althomsons SP, Hill AN, Harrist AV, et al. Statistical method to detect tuberculosis outbreaks among endemic clusters in a low-incidence setting. Emerg Infect Dis 2018;24:573–5. [PubMed: 29460749]

31. Gardy JL, Johnston JC, Ho Sui SJ, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med 2011;364:730–9. [PubMed: 21345102]

32. Guthrie JL, Delli Pizzi A, Roth D, et al. Genotyping and whole genome sequencing to identify tuberculosis transmission to pediatric patients in British Columbia, Canada, 2005–2014. J Infect Dis 2018;218:1155–63. [PubMed: 29757395]

33. Jajou R, de Neeling A, van Hunen R, et al. Epidemiological links between tuberculosis cases identified twice as efficiently by whole genome sequencing than conventional molecular typing: A population-based study. PLoS One 2018;13:e0195413.

34. Guthrie JL, Strudwick L, Roberts B, et al. Whole genome sequencing for improved understanding of Mycobacterium tuberculosis transmission in a remote circumpolar region. Epidemiol Infect 2019;147:e188.

35. Parvaresh L, Crighton T, Martinez E, Bustamante A, Chen S, Sintchenko V. Recurrence of tuberculosis in a low-incidence setting: a retrospective cross-sectional study augmented by whole genome sequencing. BMC Infect Dis 2018;18:265. [PubMed: 29879906]

36. Luo T, Yang C, Peng Y, et al. Whole-genome sequencing to detect recent transmission of Mycobacterium tuberculosis in settings with a high burden of tuberculosis. Tuberculosis (Edinb) 2014;94:434–40. [PubMed: 24888866]

37. The CRyPTIC Consortium and the 100,000 Genomes Project. Prediction of susceptibility to first-line tuberculosis drugs by DNA sequencing. N Engl J Med 2018;379:1403–15. [PubMed: 30280646]

38. Doyle RM, Burgess C, Williams R, et al. Direct whole-genome sequencing of sputum accurately identifies drug-resistant Mycobacterium tuberculosis faster than MGIT culture sequencing. J Clin Microbiol 2018;56:e00666-18.

39. Votintseva AA, Bradley P, Pankhurst L, et al. Same-Day Diagnostic and Surveillance Data for Tuberculosis via Whole-Genome Sequencing of Direct Respiratory Samples. J Clin Microbiol 2017;55:1285–98. [PubMed: 28275074]

40. Shea J, Halse TA, Lapierre P, et al. Comprehensive whole-genome sequencing and reporting of drug resistance profiles on clinical cases of Mycobacterium tuberculosis in New York State. J Clin Microbiol 2017;55:1871–82. [PubMed: 28381603]

41. Papaventsis D, Casali N, Kontsevaya I, Drobniewski F, Cirillo DM, Nikolayevskyy V. Whole genome sequencing of Mycobacterium tuberculosis for detection of drug resistance: a systematic review. Clin Microbiol Infect 2017;23:61–8. [PubMed: 27665704]

42. Dolinger DL, Colman RE, Engelthaler DM, Rodwell TC. Next-generation sequencing-based user-friendly platforms for drug-resistant tuberculosis diagnosis: A promise for the near future. Int J Mycobacteriol 2016;5 Suppl 1:S27–S8. [PubMed: 28043592]

43. Ziegler T, Mamahit A, Cox NJ. 65 years of influenza surveillance by a World Health Organization-coordinated global network. Influenza and Other Respiratory Viruses 2018;12:558–65. [PubMed: 29727518]

44. Blanton L, Dugan VG, Abd Elal AI, et al. Update: influenza activity - United States, September 30, 2018-February 2, 2019. MMWR Morb Mortal Wkly Rep 2019;68:125–34. [PubMed: 30763296]

45. Hampson A, Barr I, Cox N, et al. Improving the selection and development of influenza vaccine viruses – Report of a WHO informal consultation on improving influenza vaccine virus selection, Hong Kong SAR, China, 18–20 November 2015. Vaccine 2017;35:1104–9. [PubMed: 28131392]

46. Zhou B, Donnelly ME, Scholes DT, et al. Single-reaction genomic amplification accelerates sequencing and vaccine production for classical and Swine origin human influenza a viruses. J Virol 2009;83:10309–13. [PubMed: 19605485]

47. Zhou B, Lin X, Wang W, et al. Universal influenza B virus genomic amplification facilitates sequencing, diagnostics, and reverse genetics. J Clin Microbiol 2014;52:1330–7. [PubMed: 24501036]

48. Recommended composition of influenza virus vaccines for use in the 2019–2020 northern hemisphere influenza season. World Health Organization, 2019 (Accessed 3 September 2019, at https://www.who.int/influenza/vaccines/virus/recommendations/201902_recommendation.pdf.)

49. Addendum to the recommended composition of influenza virus vaccines for use in the 2019–2020 northern hemisphere influenza season. World Health Organization, 2019 (Accessed 3 September 2019, 2019, at https://www.who.int/influenza/vaccines/virus/recommendations/201902_recommendation_addendum.pdf.)

50. Burke SA, Trock SC. Use of influenza risk assessment tool for prepandemic preparedness. Emerg Infect Dis 2018;24:471–7. [PubMed: 29460739]

51. Cox NJ, Trock SC, Burke SA. Pandemic preparedness and the Influenza Risk Assessment Tool (IRAT). Curr Top Microbiol Immunol 2014;385:119–36. [PubMed: 25085014]

52. Russell CA, Kasson PM, Donis RO, et al. Improving pandemic influenza risk assessment. Elife 2014;3:e03883.

53. Flannery B, Zimmerman RK, Gubareva LV, et al. Enhanced genetic characterization of influenza A(H3N2) viruses and vaccine effectiveness by genetic group, 2014–2015. J Infect Dis 2016;214:1010–9. [PubMed: 27190176]

54. Uyeki TM, Katz JM, Jernigan DB. Novel influenza A viruses and pandemic threats. The Lancet 2017;389:2172–4.

55. Wilson JR, Belser JA, DaSilva J, et al. An influenza A virus (H7N9) anti-neuraminidase monoclonal antibody protects mice from morbidity without interfering with the development of protective immunity to subsequent homologous challenge. Virology 2017;511:214–21. [PubMed: 28888111]

56. Flaherty BR, Talundzic E, Barratt J, et al. Restriction enzyme digestion of host DNA enhances universal detection of parasitic pathogens in blood via targeted amplicon deep sequencing. Microbiome 2018;6:164. [PubMed: 30223888]

57. Talundzic E, Ravishankar S, Kelley J, et al. Next-generation sequencing and bioinformatics protocol for malaria drug resistance marker surveillance. Antimicrob Agents Chemother 2018;62:e02474-17.

58. Zhong D, Lo E, Wang X, et al. Multiplicity and molecular epidemiology of Plasmodium vivax and Plasmodium falciparum infections in East Africa. Malar J 2018;17:185. [PubMed: 29720181]

59. Qvarnstrom Y, Wei-Pridgeon Y, Van Roey E, et al. Purification of Cyclospora cayetanensis oocysts obtained from human stool specimens for whole genome sequencing. Gut Pathog 2018;10:45. [PubMed: 30337964]

60. Barratt JLN, Park S, Nascimento FS, et al. Genotyping genetically heterogeneous Cyclospora cayetanensis infections to complement epidemiological case linkage. Parasitology 2019:1–33.

61. Lapierre P, Nazarian E, Zhu Y, et al. Legionnaires' disease outbreak caused by endemic strain of legionella pneumophila, New York, New York, USA, 2015. Emerg Infect Dis 2017;23:1784–91. [PubMed: 29047425]

62. Levesque S, Plante PL, Mendis N, et al. Genomic characterization of a large outbreak of Legionella pneumophila serogroup 1 strains in Quebec City, 2012. PLoS One 2014;9:e103852.

63. David S, Mentasti M, Lai S, et al. Spatial structuring of a Legionella pneumophila population within the water system of a large occupational building. Microb Genom 2018;4.

64. Popovich KJ, Snitkin ES. Whole genome sequencing-implications for infection prevention and outbreak investigations. Curr Infect Dis Rep 2017;19:15. [PubMed: 28281083]

65. Snitkin ES, Zelazny AM, Thomas PJ, et al. Tracking a hospital outbreak of carbapenem-resistant Klebsiella pneumoniae with whole-genome sequencing. Sci Transl Med 2012;4:148ra16.

66. Chow NA, Gade L, Tsay SV, et al. Multiple introductions and subsequent transmission of multidrug-resistant Candida auris in the USA: a molecular epidemiological survey. Lancet Infect Dis 2018;18:1377–84. [PubMed: 30293877]

67. Oster AM, France AM, Mermin J. Molecular epidemiology and the transformation of HIV prevention. JAMA 2018;319:1657–8. [PubMed: 29630701]

68. Oster AM, France AM, Panneer N, et al. Identifying clusters of recent and rapid HIV transmission through analysis of molecular surveillance data. J Acquir Immune Defic Syndr 2018.

69. Kosakovsky Pond SL, Weaver S, Leigh Brown AJ, Wertheim JO. HIV-TRACE (TRAnsmission Cluster Engine): a tool for large scale molecular epidemiology of HIV-1 and other rapidly evolving pathogens. Mol Biol Evol 2018;35:1812–9. [PubMed: 29401317]

70. Olmstead AD, Joy JB, Montoya V, et al. A molecular phylogenetics-based approach for identifying recent hepatitis C virus transmission events. Infect Genet Evol 2015;33:101–9. [PubMed: 25917496]

71. Lockhart SR, Etienne KA, Vallabhaneni S, et al. Simultaneous emergence of multidrug-resistant Candida auris on 3 continents confirmed by whole-genome sequencing and epidemiological analyses. Clin Infect Dis 2017;64:134–40. [PubMed: 27988485]

72. Riveron JM, Ibrahim SS, Mulamba C, et al. Genome-wide transcription and functional analyses reveal heterogeneous molecular mechanisms driving pyrethroids resistance in the major malaria vector Anopheles funestus across Africa. G3 (Bethesda) 2017;7:1819–32. [PubMed: 28428243]

73. Weetman D, Wilding CS, Neafsey DE, et al. Candidate-gene based GWAS identifies reproducible DNA markers for metabolic pyrethroid resistance from standing genetic variation in East African Anopheles gambiae. Sci Rep 2018;8:2920. [PubMed: 29440767]

74. Chochua S, Metcalf BJ, Li Z, et al. Population and whole genome sequence based characterization of invasive group A streptococci recovered in the United States during 2015. MBio 2017;8:e01422–17.

75. Metcalf BJ, Chochua S, Gertz RE Jr., et al. Using whole genome sequencing to identify resistance determinants and predict antimicrobial resistance phenotypes for year 2015 invasive pneumococcal disease isolates recovered in the United States. Clin Microbiol Infect 2016;22:1002 e1- e8.

76. Topaz N, Boxrud D, Retchless AC, et al. BMScan: using whole genome similarity to rapidly and accurately identify bacterial meningitis causing species. BMC Infect Dis 2018;18:405. [PubMed: 30111301]

77. Gardy JL, Loman NJ. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet 2018;19:9–20. [PubMed: 29129921]

78. Quick J, Loman NJ, Duraffour S, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature 2016;530:228–32. [PubMed: 26840485]

79. Faria NR, Quick J, Claro IM, et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. Nature 2017;546:406–10. [PubMed: 28538727]

80. Grubaugh ND, Saraf S, Gangavarapu K, et al. Travel surveillance and genomics uncover a hidden Zika outbreak during the waning epidemic. Cell 2019;178:1057–71 e11. [PubMed: 31442400]

81. Dudas G, Carvalho LM, Bedford T, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. Nature 2017;544:309–15. [PubMed: 28405027]

82. Iwamoto M, Huang JY, Cronquist AB, et al. Bacterial enteric infections detected by culture-independent diagnostic tests--FoodNet, United States, 2012–2014. MMWR Morb Mortal Wkly Rep 2015;64:252–7. [PubMed: 25763878]

83. Marder EP, Cieslak PR, Cronquist AB, et al. Incidence and trends of infections with pathogens transmitted commonly through food and the effect of increasing use of culture-independent diagnostic tests on surveillance - Foodborne Diseases Active Surveillance Network, 10 U.S. Sites, 2013–2016. MMWR Morb Mortal Wkly Rep 2017;66:397–403. [PubMed: 28426643]

84. Huang JY, Henao OL, Griffin PM, et al. Infection with pathogens transmitted commonly through food and the effect of increasing use of culture-independent diagnostic tests on surveillance--Foodborne Diseases Active Surveillance Network, 10 U.S. Sites, 2012–2015. MMWR Morb Mortal Wkly Rep 2016;65:368–71. [PubMed: 27077946]

85. Clark SA, Doyle R, Lucidarme J, Borrow R, Breuer J. Targeted DNA enrichment and whole genome sequencing of Neisseria meningitidis directly from clinical specimens. Int J Med Microbiol 2018;308:256–62. [PubMed: 29153620]

86. Sintchenko V, Holmes EC. The role of pathogen genomics in assessing disease transmission. BMJ 2015;350:h1314.

87. Grad YH, Lipsitch M. Epidemiologic data and pathogen genome sequences: a powerful synergy for public health. Genome Biol 2014;15:538. [PubMed: 25418119]

88. Crisan A, Gardy JL, Munzner T. A systematic method for surveying data visualizations and a resulting genomic epidemiology visualization typology: GEViT. Bioinformatics 2019;35:1070–2. [PubMed: 30875428]

89. Argimon S, Abudahab K, Goater RJ, et al. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microb Genom 2016;2:e000093.

90. Hadfield J, Megill C, Bell SM, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics 2018;34:4121–3. [PubMed: 29790939]

91. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 2016;44:W242–5. [PubMed: 27095192]

92. Aldridge RW. Research and training recommendations for public health data science. Lancet Public Health 2019;4:e373.

93. Greninger AL. Societal implications of the internet of pathogens. J Clin Microbiol 2019.

94. Kostkova P. Disease surveillance data sharing for public health: the next ethical frontiers. Life Sci Soc Policy 2018;14:16. [PubMed: 29971516]

95. Metcalf BJ, Chochua S, Gertz RE Jr., et al. Short-read whole genome sequencing for determination of antimicrobial resistance mechanisms and capsular serotypes of current invasive Streptococcus agalactiae recovered in the USA. Clin Microbiol Infect 2017;23:574 e7- e14.

96. Grad YH, Harris SR, Kirkcaldy RD, et al. Genomic epidemiology of gonococcal resistance to extended-spectrum cephalosporins, macrolides, and fluoroquinolones in the United States, 2000–2013. J Infect Dis 2016;214:1579–87. [PubMed: 27638945]

97. Glebova O, Knyazev S, Melnyk A, et al. Inference of genetic relatedness between viral quasispecies from sequencing data. BMC Genomics 2017;18:918. [PubMed: 29244009]

98. Longmire AG, Sims S, Rytsareva I, et al. GHOST: global hepatitis outbreak and surveillance technology. BMC Genomics 2017;18:916. [PubMed: 29244005]

Box:

Generalizations about Sequencing in Public Health

Several attributes of next-generation sequencing are driving adoption of the technology within public health:

- High resolution subtyping of pathogens
  - Examples
    - ♦ Bacterial enteric illness: improves detection of and response to outbreaks.
    - ♦ Tuberculosis: allows better targeting of interventions to stop transmission.
    - ♦ *Legionella:* provides new tool to understand the ecology of the pathogen in water systems.
    - ♦ Potential agents of bioterrorism: allows for improved forensics.
  - *Caveats:* Legacy technologies often need to be continued during a transition period, since older subtyping characterizations often can’t be reliably predicted from nucleic acid sequence data; PFGE patterns, for example, usually cannot be predicted from routine whole-genome sequencing.
- Efficient inference of phenotypic traits
  - Examples
    - ♦ Serotyping: In US public health labs, influenza viruses are now subject to a “sequencing first” approach, in which antigenic type and subtype can be inferred from the sequence; only a subset of viruses undergo traditional typing and subtyping. For pathogens such as *E. coli,*[15,18,24] *Salmonella*[25,26] or pneumococcus,[95] serotype can usually be inferred from sequence data, without the need to acquire and maintain serum panels.
    - ♦ Antimicrobial resistance: for bacterial pathogens such as Salmonella,[21,22] Escherichia coli,[23] Streptococcus,[75,95] MTB, or gonococcus[96] (to name but a few), antimicrobial resistance is increasingly inferred from genomic data.
    - ♦ Virulence: known virulence factors, such as the presence or absence of Shiga toxin genes in an *E. coli* strain, can also be inferred from genomic data.
  - *Caveats:* There will probably always be a need for traditional phenotyping. The ability to predict a phenotype from a genome

generally relies on known correlations between the phenotypic characteristics and specific genetic sequences. Particularly in rapidly evolving species such as influenza, those correlations will need constant updating. In addition, the consistency of those correlations is variable. The reliability of inferred antimicrobial resistance, for example, is highly dependent on the type of antibiotic, the mechanism of resistance, and the species of bacteria. This reliability should improve over time as more data become available and algorithms for predicting phenotype improve. The capability to infer phenotype from genotype means that fewer traditional tests will need to be done in the future, and that fewer laboratories (i.e., reference laboratories) will need to maintain the capacity to perform them.

- *"Deep sequencing" rather than "consensus sequencing".* Whereas Sanger sequencing generally provides a single, "consensus" sequence from a sample, NGS typically provides many (often hundreds, thousands or more) "reads" of the gene or amplicon.
  - Examples
    - Malaria: In highly endemic areas, infection with multiple strains of malaria is common. In such cases, Sanger sequencing usually reflects only the most dominant strain in the individual and can miss the presence of other strains, which may have differing resistance to anti-malarial agents. In malaria endemic areas, deep sequencing can also be used to quantify the number of strains in an individual, a correlate of the intensity of transmission and potentially a tool for evaluating the impact of community interventions.
    - Hepatitis C: Hepatitis C virus mutates rapidly in individuals, resulting in a "swarm of quasispecies". Data on the diversity of quasispecies in two individuals provides a reliable means of inferring whether they are part of a single outbreak.[97,98]
    - Influenza: High mutation rates in influenza virus can also lead to minor variants with resistance to oseltamivir or other antivirals, which could be missed by consensus sequencing.
  - *Caveats:* Sequencing errors, which are more common with NGS than with Sanger sequencing, can create the illusion of rare variants. Careful analysis of potential sequence variants is needed to prevent this.

- *More efficient workflows.* Characterizing pathogens by means of sequencing is sometimes but not universally less laborious and less expensive than traditional typing.
  - Examples
    - *Escherichia coli:* Shiga toxin-producing *E coli* (STEC) are a common cause of foodborne illness. A standard workflow for characterizing STEC involves determination of serotype, traditionally accomplished by means of a panel of antisera to determine O antigen and, if indicated, H antigen, as well as detection of Shiga toxin or Shiga toxin genes by, for example, PCR. All of these characteristics, as well as susceptibility or resistance to several antibiotics, can be reliably inferred from whole-genome sequencing.
    - Influenza virus: The United States is now using a “sequence first” approach to influenza virus characterization (see text).
- *Caveats:* Sequencing, for the time being, is often more expensive than traditional subtyping alone. For bacterial pathogens, for example, whole-genome sequencing is typically twice as expensive as PFGE alone. However, if PFGE for a particular pathogen needs to be accompanied by traditional phenotyping such as serotyping, virulence typing, or antimicrobial resistance testing, and those features can be reliably inferred from the genome, then the sequencing approach is often less expensive.

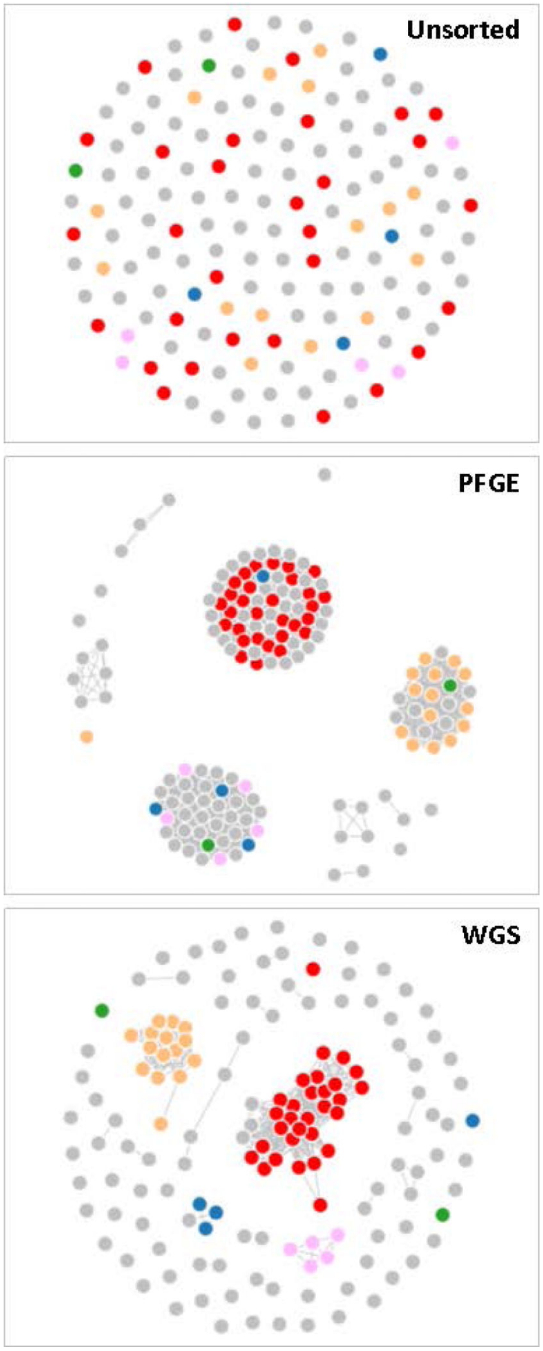

Figure 1: Example of sequencing for outbreak detection and investigation.

An important purpose of infectious disease surveillance is identifying 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 by detecting 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 *Salmonella* Enteritidis gastroenteritis. 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 (WGS) linked these two clusters together and to several other cases outside the region.

The first panel (“Unsorted”), displays cases randomly, without regard to molecular subtyping.

The second panel represents a grouping of cases based on results of pulsed-field gel electrophoresis (PFGE), a molecular subtyping technology that US public health agencies have used since the 1990s. In this example, PFGE was mostly successful at grouping cases from the largest (red) outbreak; however, the group includes many cases unrelated to the outbreak, complicating the investigation and reducing the likelihood of finding the food source.

In the third panel, we show that the finer resolution afforded by whole-genome sequencing (WGS) 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 case, epidemiologic investigation identified shell eggs as the likely source, which was quickly confirmed by isolating *Salmonella* Enteritidis from the implicated eggs and confirming that its WGS 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 (green) were linked to live poultry exposure as part of a much larger, multistate, multi-strain outbreak that occurred mostly outside the region shown here. The five cases in light pink were investigated as an outbreak but no food source was identified. The fifteen cases in light orange occurred in a state where real-time WGS had not yet been implemented; their isolates were not sequenced until a later date, after the apparent outbreak had ended. This figure summarizes relationships identified by WGS using 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.

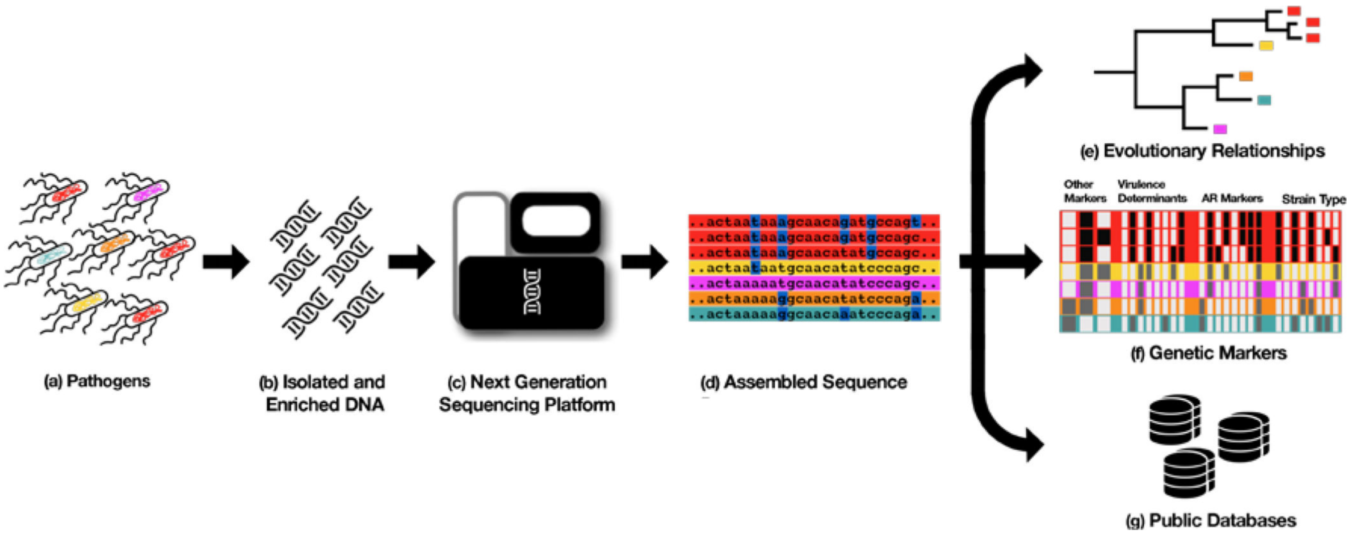
(e) Evolutionary Relationships
Other Markers
Virulence Determinants
AR Markers
Strain Type
(f) Genetic Markers
..actaataaagcaacagatgccagt..
..actaataaagcaacagatgccagc..
..actaataaagcaacatatgccagc..
..actaataatgcaacatatcccagc..
..actaaaaatgcaacatatcccagc..
..actaaaaaggcaacatatcccaga..
..actaaaaaggcaacaaatcccaga..
(a) Pathogens
(b) Isolated and Enriched DNA
(c) Next Generation Sequencing Platform
(d) Assembled Sequence
(g) Public Databases

Figure 2.

Typical Pathogen Genomics Workflow. From pathogens (a) collected in the course of disease surveillance, genomic DNA (b) or RNA are extracted, shorn into shorter segments, labeled, and subjected to next-generation (high-throughput) sequencing (c). The raw data from the sequencer are sorted, reassembled and aligned to other genomes for comparison (d). The assembled genomes are used for several purposes including determining relatedness (e) and predicting phenotypic traits such as virulence, antimicrobial resistance, and serotype (f). Increasingly, the data made publicly available in real-time for use by researchers and for the development of diagnostics, therapeutics and vaccines (g).