AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Clinical Microbiology®

MINIREVIEW

Real-Time Analysis and Visualization of Pathogen Sequence Data

Richard A. Neher,[a,b] Trevor Bedford[c]

[a]Biozentrum, University of Basel, Basel, Switzerland
[b]SIB Swiss Institute of Bioinformatics, Basel, Switzerland
[c]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

ABSTRACT The rapid development of sequencing technologies has to led to an explosion of pathogen sequence data, which are increasingly collected as part of routine surveillance or clinical diagnostics. In public health, sequence data are used to reconstruct the evolution of pathogens, to anticipate future spread, and to target interventions. In clinical settings, whole-genome sequencing can identify pathogens at the strain level, can be used to predict phenotypes such as drug resistance and virulence, and can inform treatment by linking closely related cases. While sequencing has become cheaper, the analysis of sequence data has become an important bottleneck. Deriving interpretable and actionable results for a large variety of pathogens, each with its own complexity, from continuously updated data is a daunting task that requires flexible bioinformatic workflows and dissemination platforms. Here, we review recent developments in real-time analyses of pathogen sequence data, with a particular focus on the visualization and integration of sequence and phenotype data.

KEYWORDS bioinformatics, molecular epidemiology, phylogenetic analysis

As pathogens replicate and spread, their genomes accumulate mutations. These changes can now be detected via inexpensive and rapid whole-genome sequencing (WGS) on an unprecedented scale. Such sequence data are increasingly being used to track the spread of pathogens and to predict their phenotypic properties. Both applications have great potential to inform public health and treatment decisions if sequence data can be obtained and analyzed rapidly. Historically, however, sequencing and analysis have lagged months to years behind sample collection. The results from those studies have taught us much about pathogen molecular evolution, genotype-phenotype maps, and epidemic spread but almost always have come too late to inform public health interventions or treatment decisions.

The rapid development of sequencing technologies has made routine sequencing of viral and bacterial genomes possible, and tens of thousands of whole-genome sequences are deposited in databases every year (Fig. 1). Many more genomes are sequenced and, regrettably, not shared. There are currently two major ways in which high-throughput sequencing technologies are used in public health and diagnostic applications, (i) to track outbreaks and epidemics to inform public health responses and (ii) to characterize individual infections to tailor treatment decisions.

Sequencing in public health. The utility of rapid sequencing and phylogenetic analysis of pathogens is perhaps most evident for influenza viruses and foodborne diseases. Due to rapid evolution of their viral surface proteins, the antigenic properties of the circulating influenza viruses change every few years, and the seasonal influenza vaccine needs frequent updating (1). The WHO Global Influenza Surveillance and Response System (GISRS) sequences hundreds of viruses every month, and many of

Accepted manuscript posted online 22 August 2018

Citation Neher RA, Bedford T. 2018. Real-time analysis and visualization of pathogen sequence data. J Clin Microbiol 56:e00480-18. https://doi.org/10.1128/JCM.00480-18.

Editor Colleen Suzanne Kraft, Emory University

Address correspondence to Richard A. Neher, richard.neher@unibas.ch.

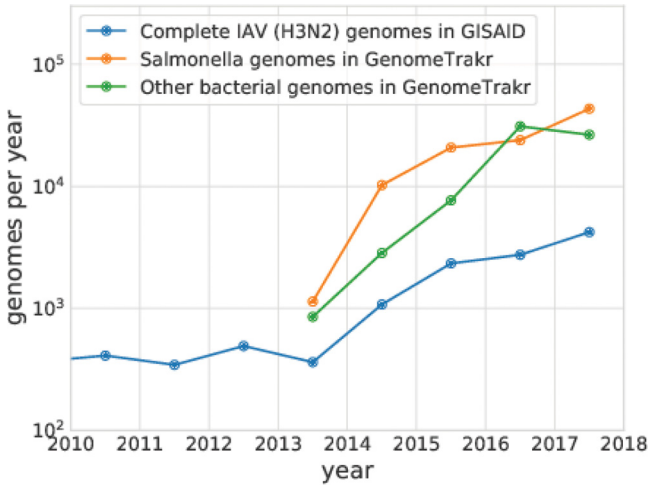

FIG 1 Numbers of complete pathogen genomes, which have increased dramatically over the past few years. More than 4,000 complete influenza A (IAV) subtype H3N2 virus genomes were deposited in GISAID in 2017. The GenomeTrakr network sequenced more than 40,000 *Salmonella* genomes and 25,000 other bacterial genomes (mostly *Listeria*, *Escherichia coli/Shigella*, and *Campylobacter*) in 2017 (11).

these sequences are submitted to the GISAID database (https://www.gisaid.org) within 4 weeks after sample collection. Phylogenetic analysis of these data provides an accurate and up-to-date summary of the spread and abundance of different viral variants, which is crucial input for the biannual consultations on seasonal influenza vaccine composition.

Such rapid turnaround and data sharing is considerably harder to achieve in an outbreak setting under resource-limited conditions. However, Quick et al. (2) achieved even shorter turnaround times during the end of the West African Ebola outbreak in 2014 to 2015. Similarly, Dyrdak et al. (3) analyzed an enterovirus outbreak in Sweden and continuously updated the manuscript until publication, with sequences sampled within days of publication being included in the analysis.

Molecular epidemiological techniques can reconstruct the temporal and spatial spread of an outbreak. In this case, the accumulation of mutations alongside a molecular clock estimate can be used to date the origin of an outbreak. Similarly, by linking samples that originate from different geographic locations, phylogeographic methods can reconstruct the geographic spread and can differentiate distinct introductions. The resolution of these inferences critically depends on the rate at which mutations accumulate in the sequenced locus, which increases with the per-site evolutionary rate and the length (L) of the locus.

RNA viruses accumulate changes in their genomes with a typical rate of 0.0005 to 0.005 changes per site per year (4). Rate estimates vary from virus to virus and depend on the time scale of observation and whether changes are measured within or between hosts. Ebola virus and Zika virus, for example, evolve at a rate μ of ~0.001 changes per site per year. The expected time interval without a substitution along a transmission chain is $1/(\mu L)$, which corresponds to approximately 5 weeks for Zika virus (L of ~10 kb) and 3 weeks for Ebola virus (L of ~19 kb). Hence, the evolution and spread of such RNA viruses can be resolved on the scale of 1 month. While this temporal resolution is typically insufficient to resolve individual transmissions, it is high compared to the duration of outbreaks. Therefore, rapid sequencing and analysis have the potential to inform intervention efforts as outbreaks are unfolding. In particular, they can rule out direct transmission and differentiate different introductions or zoonosis.

Phylodynamic and phylogeographic methods are best established for viral pathogens with high evolutionary rates and small genomes, for which large-scale sequencing has been possible for years. The evolutionary rates of bacteria are many orders of magnitude lower than those of RNA viruses. But bacteria also have about 100- to

1,000-fold larger genomes, and it is now possible to sequence entire bacterial genomes at low cost. Substitution rate estimates for bacteria involve substantial uncertainty, but they tend to be on the order of 1 substitution per megabase per year (with variations of about 1 to 2 orders of magnitude between species [5]). With a typical genome size of 5 Mb, these rates translate into 5 to 10 substitutions per genome per year, similar to genome-wide substitution rates of many RNA viruses. The substitution rate in the core genome of methicillin-resistant *Staphylococcus aureus* (MRSA), for example, was estimated to be 1.3×10^{-6} changes per site per year (6). The core genome of *Listeria monocytogenes* evolves more slowly, at about 1 substitution every 2.5 y (7). Hence, real-time phylogenetic analysis for bacterial outbreak tracking is possible in much the same way as it is for RNA viruses. Analysis of bacterial genomes, however, is vastly more complicated than is that of RNA viruses with short genomes. Bacteria frequently exchange genetic material via horizontal transfer, take up genes from the environment, and rearrange their genomes. Recombination can blur phylogenetic signals, and recombinant sequences are often difficult to remove. Furthermore, strong selection within hosts, for example through drug therapy, can accelerate evolution by up to 1 order of magnitude (8). If not properly accounted for, these processes can blur any temporal signal and obscure links between closely related isolates.

Even with whole genomes, phylogenetic resolution typically is insufficient to make the case for direct transmission, but transmission can be confidently ruled out for divergent sequences, seemingly unrelated cases can be grouped into outbreaks (e.g., an outbreak of drug-resistant *Mycobacterium tuberculosis* among migrants arriving in multiple European countries [9]), and predominant routes of transmission and likely sources in the environment or animal reservoirs can be identified. GenomeTrakr and PulseNet, for example, represent large federated efforts to sequence tens of thousands of genomes from foodborne outbreaks and clinical samples (10, 11). All sequence data from these projects are publicly available in NCBI databases with little delay and are analyzed in real time to track outbreaks. The recently released NCBI Pathogen Detection system (https://www.ncbi.nlm.nih.gov/pathogens) provides convenient access to the sequence data and metadata generated by these projects, as well as phylogenetic analysis.

These examples illustrate the potential and feasibility of obtaining actionable information from pathogen sequence data for both viral and bacterial infections. With rapidly increasing data volumes, however, efficient processing pipelines and tools that help with interpretation (e.g., visualizations) increasingly become the bottleneck.

Sequencing in diagnostics and therapy. For some pathogens, such as Zika virus, sequencing of the genome has no implications for treatment. In the case of HIV, however, drug resistance profiles derived from sequence data have directly informed treatment for years (12). As the genetic basis of drug resistance phenotypes becomes better understood, rapid WGS will increasingly be used to diagnose and to phenotype pathogens directly from clinical specimens. Such culture-free methods are particularly important for tuberculosis, for which culture-based susceptibility testing takes many weeks. Votintseva et al. (13) recently showed that high-throughput sequencing directly from respiratory samples could provide drug resistance profiles for *M. tuberculosis* within 1 day.

Sequencing for diagnostic purposes and sequencing for public health surveillance have different aims and requirements but can complement each other. Public health responses typically require recent data with an emphasis on dynamics. Surveillance data provide context for individual cases, and clinical treatment requires a stable database with validated content to make reliable predictions regarding drug susceptibility, phylogenetic context, and protective measures. Clinical sequence data should be entered into surveillance databases immediately, whenever ethically possible. Only with rapid open sharing of sequence data can the full potential of molecular epidemiology be realized (11).

The challenges involved in sample collection, processing, sequencing, and data sharing have been discussed at length elsewhere (14). Here, we focus on software developments that facilitate the implementation of real-time analysis, with an emphasis on web-based visualization, as a full review of general tools for genomic analysis and visualization is not easily accomplished.

RAPID AND INTERPRETABLE ANALYSIS OF GENOMIC DATA

A typical molecular epidemiological analysis aims to identify transmission clusters, to date the introduction of the pathogen, to detail the geographic spread, and in some cases to identify potential phenotypic changes of a pathogen from sequence data. The rapidly increasing numbers of sequenced genomes make comprehensive analysis computationally challenging. While thousands of viral genomes can be aligned within minutes (e.g., by MAFFT) and the reconstruction of a basic phylogenetic tree typically takes less than 1 h (e.g., using IQ-TREE, RAxML, or FastTree), the most popular tool for phylodynamic inference (BEAST) (15) often takes weeks to finish.

To overcome these hurdles, several tools that use simpler heuristics have been developed to infer time-stamped phylogenies (16–18). Rather than sampling a large number of tree topologies, these tools use the topology of an input tree with little or no modification. Dating of ancestral events tends to be of comparable accuracy, compared to BEAST (16–18). However, these tools do not integrate the uncertainty of tree reconstruction and provide limited flexibility to infer demographic models. Furthermore, the heuristics used by these programs are based on assumptions (for example, that sequences are closely related), and they are not expected to be accurate in all scenarios. The computational cost of Bayesian phylodynamics could be mitigated if methods for continuous updating and augmenting of the Markov chain with additional data were developed. For the present time, however, efficient heuristics and sensible approximations deliver sufficiently accurate and reliable results when near-real-time analysis is required.

Nextflu and Nextstrain for viral genomes. The number of influenza viruses that are sequenced and phenotyped per month has increased sharply, to the point that comprehensive and timely manual analysis and annotation of the results is no longer feasible. In 2014, we developed an automated phylodynamic analysis pipeline that operates on an up-to-date database of sequences and serological information. The results of this pipeline were made available at nextflu.org and included a phylogeny, branch-specific mutations, frequency trajectories of mutations and variants, and a model of antigenic evolution.

Nextflu is now integrated into the more general platform Nextstrain, which provides an online platform for outbreak investigations of diverse viruses and is available at https://nextstrain.org (19). Nextstrain uses TreeTime (18) to infer time-scaled phylogenies and to conduct ancestral sequence inference. In addition, Nextstrain uses the discrete ancestral character inference of TreeTime to infer the likely geographic state of ancestral nodes. Since this approach applies mutation models to migration, it is often called a "mugration" model. A phylodynamic/phylogeographic analysis of 1,000 sequences of 10-kb length takes on the order of 1 hour on a standard laptop computer.

Bacterial WGS data. Bacterial WGS data typically come in the form of millions of short reads, which can be assembled into contigs, mapped against reference sequences, or classified based on kmer distributions. A large number of tools have been developed for rapid species identification, typing, and variant calling. Pathogenwatch (formerly known as WGSA), for example, allows users to upload an assembly and can detect the species and infer the multilocus sequence type within a few seconds. In addition, Pathogenwatch predicts antibiotic resistance profiles for a number of species. Pathogenwatch was developed by the Center for Genomic Pathogen Surveillance and is available at https://pathogen.watch.

Bacterial genomes are very dynamic and frequently gain or lose genes. Even closely related bacteria can differ in the presence or absence of dozens of genes. To track transmission and to detect clusters, genomes are typically compared at a set of core

genes that are present in all bacteria of a species. Genes present in only a fraction of individuals are referred to as accessory genes.

Clinically important genes such antibiotic resistance determinants and virulence factors are often not part of the core genome and are horizontally transferred among strains and species. Therefore, collections of bacterial genomes are analyzed using pan-genome tools that aim to cluster all genes in the collection of genomes into orthologous groups. Early methods for pan-genome analysis scaled poorly with the number of genomes analyzed, since every gene in every genome needed to be compared to every other gene. The first tool capable of analyzing hundreds of bacterial genomes was Roary (20). Roary is designed to work with very similar genomes (as is common in outbreak scenarios) and accelerates the inference of orthologous gene clusters by preclustering genomes. A more recent pan-genome analysis pipeline capable of large-scale analysis is panX (21), which speeds up clustering by hierarchically building up the complete pan-genome from sub-pan-genomes inferred from smaller batches of genomes. PanX is coupled to a web-based visualization platform, as discussed below.

While the pan-genome tools cluster annotated genes in the collection of genomes, they are of little help in assessing the origin and distribution of a particular sequence. Traditional tools for homology searches in NCBI databases index only assembled sequences, but today the majority of sequence data are stored in short-read archives rather than GenBank. Bradley et al. (22) developed a method to search the entire collection of microbial sequence data, including metagenomic samples from a wide variety of environments. The ability to search this vast treasure trove of data will likely be transformative for assessing the spread and prevalence of novel resistance determinants. The recently discovered mobile colistin resistance gene *mcr-1*, for example, was found in more than 100 data sets in which it had not been described previously (22).

Outlook. Most current analysis pipelines require rerunning the entire analysis when even a single sequence is added. While this strategy is still feasible today, it will likely become unsustainable in the future. Applications that support inexpensive updating of data sets and online additions of user data will likely replace current versions.

VISUALIZATION AND INTERPRETATION

With increasing data set sizes, interpretation and exploration of data become increasingly challenging. Phylogenetic trees can be visualized as familiar planar graphs, but the trees alone show only genetic similarity between isolates and quickly become unintelligible as the number of sequences increases. For pathogen sequence data to be truly useful, the data need to be integrated with other types of information, ideally in an interactive way. A suitable platform to do so is the web browser, and several powerful web applications have emerged in the past few years. In addition, browser-based visualizations are naturally disseminated online.

Microreact. Microreact is a web application based on React (a JavaScript framework for interactive applications), D3.js (a JavaScript library for producing dynamic interactive data visualizations), Phylocanvas (a JavaScript flexible tree viewer), and Leaflet (a JavaScript mapping toolkit) (23). Microreact allows exploration of a phylogenetic tree, the geographic locations, and a timeline of the samples; it is available at https://microreact.org. Custom data sets can be loaded into the application in the form of a Newick tree and a tabular file containing information such as geographic locations or sampling data.

Nextstrain. Nextstrain was developed as a more generic and more flexible version of Nextflu (19) and is available at https://nextstrain.org. Similar to Microreact, Nextstrain uses React, D3.js, and Leaflet, but it uses a custom tree viewer that has flexible zooming and annotation options. The tree can be decorated with any discrete or continuous attribute (for example, geographic locations), both on the tips of the tree and, with inferred values, on internal nodes. Nextstrain maps individual mutations to branches in the tree and thereby allows mutations to be associated with phenotypes or geographic

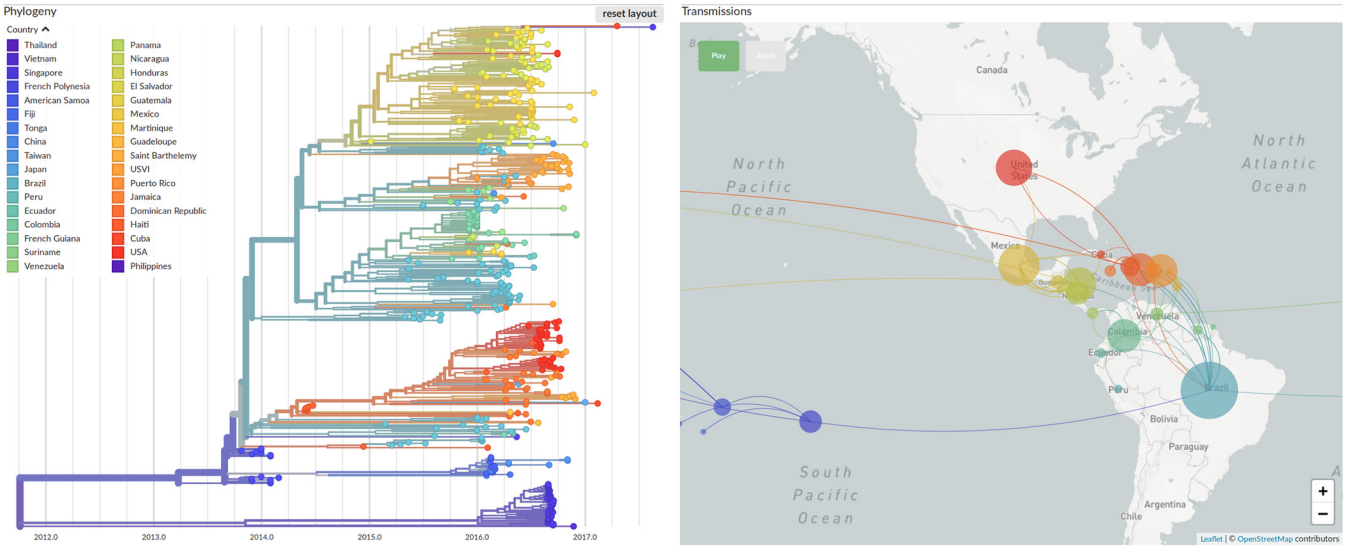

FIG 2 Phylogeographic analysis of Zika virus sequences using https://nextstrain.org (19). Whole-genome sequences sampled between 2013 and 2017 were processed using the Nextstrain pipeline. Nextstrain reconstructs the likely time and place of each internal node of the tree and from this assignment infers possible transmission patterns, which are displayed on a map. Molecular analysis of this sort reveals, for example, multiple introductions of Zika virus into Florida, most likely originating from viruses circulating in the Caribbean region in 2015 and 2016. Map by OpenStreetMap contributors (https://www.openstreetmap.org/copyright).

distributions. The map in Nextstrain shows putative transmission events, and a panel indicates genetic diversity across the genome (Fig. 2).

The analyses by Nextstrain and Nextflu critically depend on timely and open sharing of sequence information, which many laboratories around the globe contribute. To incentivize early prepublication sharing of data, platforms like Nextstrain need to explicitly acknowledge individual contributions. Ideally, such platforms should provide added value to authors, such as deep links that show data contributed by a particular group in the context of the outbreak.

Phandango. Phandango is an interactive viewer for bacterial WGS data (24) that combines a phylogenetic tree with metadata columns and gene presence/absence maps or recombination events. Phandango is available at https://phandango.net and can ingest the output of a number of analysis tools commonly used for the analysis of bacterial WGS data, such Gubbins, Roary, and BRAT.

PanX. PanX is a pan-genome analysis pipeline that is coupled to web-browser-based visualization (21). Similar to Phandango, it displays a core genome single-nucleotide polymorphism (SNP) phylogeny, but it is more centered on genetic variations in individual genes. Pan-genomes of about 100 bacterial species, based on curated reference genomes, are available at pangenome.de. The tree and alignment of each gene in the pan-genome can be accessed rapidly by searching a table of gene names and annotations. PanX then displays gene and species trees side by side and maps gene gain and loss events to branches in the core genome tree and mutations to branches in the gene tree. Trees can be colored according to arbitrary attributes, such as resistance phenotypes, and associations between genetic variations and the phenotypes can be explored.

Other tools. SpreaD3 allows visualization of phylogeographic reconstructions from models implemented in the software package BEAST (25). PhyloGeoTool is a web application to navigate large phylogenies interactively and to explore associated clinical and epidemiological data (26). TreeLink displays phylogenetic trees alongside metadata in an interactive web application (27).

CHALLENGES IN DATA INTEGRATION AND VISUALIZATION

With rapidly increasing volumes of sequence data, decisions regarding how the data are filtered and what analyses are shown become increasingly important. Epidemiological investigations of a novel outbreak typically seek to identify the sources, to track

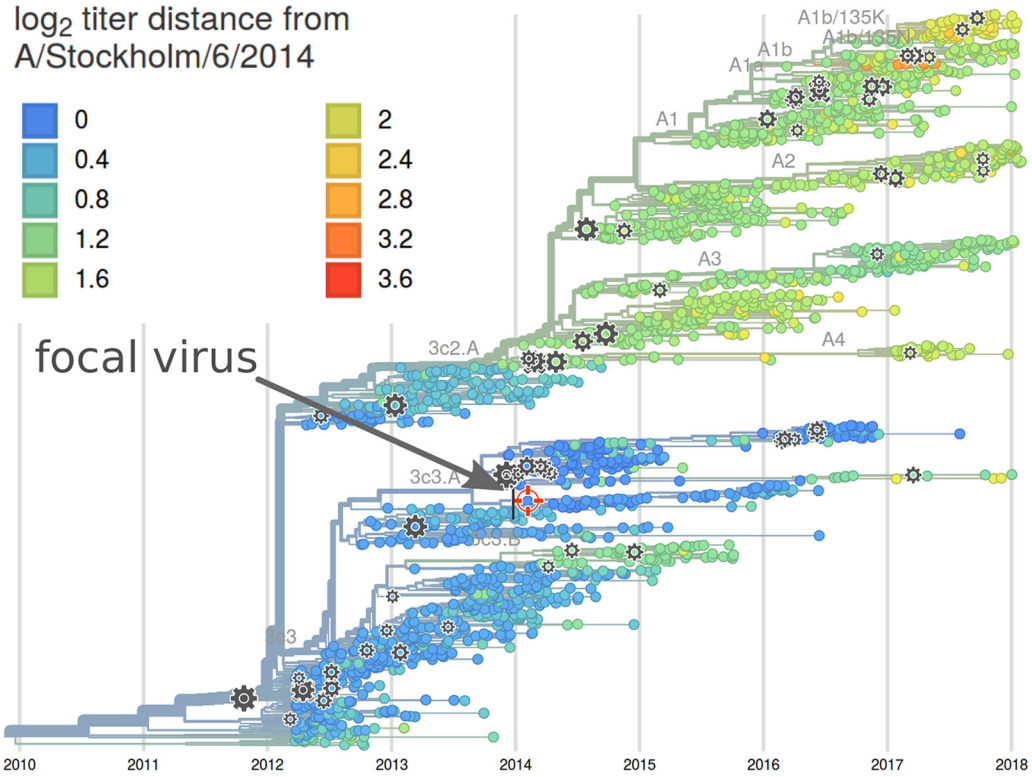

FIG 3 Integration of HI titer data with molecular evolution of the influenza virus. Each year, influenza laboratories determine thousands of HI titers for test viruses, relative to sera raised against several reference viruses (indicated by gray cogs). These data can be integrated with the molecular evolution of the virus and visualized in the phylogeny (here showing titers inferred using a model). The reference virus with respect to which titers are displayed can be chosen by clicking on the corresponding symbol in the tree (29). The visualization presents both raw data (via tooltips for each virus) and a model inference that integrates many individual measurements (hi.nextflu.org).

the spread, and to detect transmission chains. In such cases, a generic combination of map, tree, and timeline often is an appropriate and sufficient visualization. Nextstrain and Microreact both follow this paradigm.

In analyses of established pathogens that continuously adapt to treatments, vaccines, or preexisting immunity, however, more specific applications are necessary, since case data, phenotype data, and clinical parameters differ wildly by pathogen. Such data generally have a common core (such as sample date and location), but other parameters (such as drug resistance phenotype, disease severity, host age, risk group, and serology data) are pathogen specific. These data are at least as important as phylodynamic inferences from sequence data for interpretation of the epidemiological dynamics, but ethical and technical considerations can impede data sharing. The value of both types of data is greatly increased by seamless integration, but the idiosyncrasies require flexible analysis and visualization frameworks that can be tailored to specific pathogens.

One such example is the serological characterization of influenza viruses via hemagglutination inhibition (HI) titers, using antisera raised in ferrets. Such titers, which are routinely recorded by the GISRS to monitor the antigenic evolution of influenza viruses, are a good example of how phenotype information can be interactively integrated with information on phylogeny and molecular evolution. HI titers are reported in large tables and traditionally have been visualized using multidimensional scaling, without any reference to the phylogeny. In studies reported by Bedford et al. (28) and Neher et al. (29), we developed methods to integrate data on the molecular and antigenic evolution of influenza viruses. This integration allows association of genotypic changes with antigenic evolution and suggests intuitive and interactive visualization of HI titer data on the phylogeny. A screenshot of this integration is shown in Fig. 3. Most HI titer data are not

openly available, due to data-sharing restrictions, but historical data from McCauley and colleagues can be visualized with the molecular evolution at hi.nextflu.org.

In addition to phenotype integration, it is crucial to choose the right level of detail for a specific application. This is particularly true for bacteria, for which the relevant information might be the core genome phylogeny, the presence/absence of particular genes or plasmids, or individual mutations in specific genes. If the analysis tool and the visualization do not provide a fine-grained analysis at the relevant level, then the most important patterns might remain hidden. Conversely, sifting through every gene or mutation is prohibitive. The primary aim should be to highlight the most important and robust patterns and to provide flexible methods to filter and to rank variants (e.g., by recent increases in frequency or associations with the host, resistance, or risk group). Users should have the possibility to expose detail on demand when deeper exploration is required.

Similarly, parameter inferences and model abstractions are very useful for obtaining a concise summary of the data, but they should be complemented by the ability to interrogate the raw data (e.g., an estimate of the evolutionary rate should be accompanied by a scatter plot of root-to-tip divergence and sampling times). This is particularly important in outbreak scenarios, when methods are being applied to an emerging pathogen in a developing situation. For clinical applications, the presentation of the results of an analysis should be focused on the sample in question and should provide only reliable and actionable information; suggestive and correlative results tend to be a distraction (30).

CONCLUSIONS

High-throughput and rapid sequencing is revolutionizing diagnostic and epidemiological analyses of infectious diseases. Sequence data can be used to identify pathogens unambiguously, to link related cases, and to reconstruct the spread of an outbreak and will soon allow detailed prediction of a pathogen's phenotype. The GISRS is a good example of a near-real-time surveillance system. Hundreds of viruses are sequenced and phenotyped every month, and the sequence data are shared in a timely manner. A global comprehensive analysis of these data, updated about once a week, is available at https://nextstrain.org/flu. These analyses directly inform the selection process for influenza vaccine strains (1).

Several public health agencies have adopted WGS as their primary tool for outbreak investigation, and many centers share the data openly with commendable timeliness. The GenomeTrakr and PulseNet networks, for example, now sequence and openly release about 5,000 bacterial genomes per month (10, 11). These data are accessible through the recently released NCBI Pathogen Detection system at https://www.ncbi.nlm.nih.gov/pathogens, with analysis results being available via FTP transfer.

These two examples clearly show that near-real-time genomic surveillance is possible and valuable and all of the individual components to implement such surveillance are in place. To realize this potential for many more pathogens, however, sample collection and sequencing need to be streamlined, data need to be shared along with the relevant metadata, and specific analysis methods and visualizations need to be implemented and maintained.

ACKNOWLEDGMENTS

We are indebted to scientists around the globe for timely and open sharing of sequence data, without which real-time visualization would not be worthwhile. We are grateful to James Hadfield for critical reading of the manuscript and to Greg Armstrong for valuable feedback and insights into WGS surveillance efforts for bacteria.

T.B. is a Pew Biomedical Scholar and is supported by NIH R35 GM119774-01.

REFERENCES

1. Morris DH, Gostic KM, Pompei S, Bedford T, Łuksza M, Neher RA, Grenfell BT, Lassig M, McCauley JW. 2018. Predictive modeling of influenza shows the promise of applied evolutionary biology. Trends Microbiol 26:102–118. https://doi.org/10.1016/j.tim.2017.09.004.

2. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, Bore JA, Koundouno R, Dudas G, Mikhail A, Ouédraogo N, Afrough B, Bah A, Baum JH, Becker-Ziaja B, Boettcher JP, Cabeza-Cabrerizo M, Camino-Sanchez A, Carter LL, Doerrbecker J, Enkirch T, Dorival IGG, Hetzelt N,

Hinzmann J, Holm T, Kafetzopoulou LE, Koropogui M, Kosgey A, Kuisma E, Logue CH, Mazzarelli A, Meisel S, Mertens M, Michel J, Ngabo D, Nitzsche K, Pallash E, Patrono LV, Portmann J, Repits JG, Rickett NY, Sachse A, Singethan K, Vitoriano I, Yemanaberhan RL, Zekeng EG, Trina R, Bello A, Sall AA, Faye O, Faye O, Magassouba N, Williams CV, Amburgey V, Winona L, Davis E, Gerlach J, Washington F, Monteil V, Jourdain M, Bererd M, Camara A, Somlare H, Camara A, Gerard M, Bado G, Baillet B, Delaune D, Nebie KY, Diarra A, Savane Y, Pallawo RB, Gutierrez GJ, Milhano N, Roger I, Williams CJ, Yattara F, Lewandowski K, Taylor J, Rachwal P, Turner D, Pollakis G, Hiscox JA, Matthews DA, O'Shea MK, Johnston AM, Wilson D, Hutley E, Smit E, Di Caro A, Woelfel R, Stoecker K, Fleischmann E, Gabriel M, Weller SA, Koivogui L, Diallo B, Keita S, Rambaut A, Formenty P, Gunther S, Carroll MW. 2016. Real-time, portable genome sequencing for Ebola surveillance. Nature 530:228–232. https://doi.org/10.1038/nature16996.

3. Dyrdak R, Grabbe M, Hammas B, Ekwall J, Hansson KE, Luthander J, Naucler P, Reinius H, Rotzen-Ostlund M, Albert J. 2016. Outbreak of enterovirus D68 of the new B3 lineage in Stockholm, Sweden, August to September 2016. Euro Surveill 21:30403. https://doi.org/10.2807/1560-7917.ES.2016.21.46.30403.
4. Duffy S, Shackelton LA, Holmes EC. 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat Rev Genet 9:267–276. https://doi.org/10.1038/nrg2323.
5. Duchene S, Holt KE, Weill FX, Le Hello S, Hawkey J, Edwards DJ, Fourment M, Holmes EC. 2016. Genome-scale rates of evolutionary change in bacteria. Microb Genom 2:e000094.
6. Holden MT, Hsu LY, Kurt K, Weinert LA, Mather AE, Harris SR, Strommenger B, Layer F, Witte W, de Lencastre H, Skov R, Westh H, Zemlicková H, Coombs G, Kearns AM, Hill RL, Edgeworth J, Gould I, Gant V, Cooke J, Edwards GF, McAdam PR, Templeton KE, McCann A, Zhou Z, Castillo-Ramírez S, Feil EJ, Hudson LO, Enright MC, Balloux F, Aanensen DM, Spratt BG, Fitzgerald JR, Parkhill J, Achtman M, Bentley SD, Nübel U. 2013. A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant *Staphylococcus aureus* pandemic. Genome Res 23:653–664. https://doi.org/10.1101/gr.147710.112.
7. Moura A, Criscuolo A, Pouseele H, Maury MM, Leclercq A, Tarr C, Bjrkman JT, Dallman T, Reimer A, Enouf V, Larsonneur E, Carleton H, Bracq-Dieye H, Katz LS, Jones L, Touchon M, Tourdjman M, Walker M, Stroika S, Cantinelli T, Chenal-Francisque V, Kucerova Z, Rocha EPC, Nadon C, Grant K, Nielsen EM, Pot B, Gerner-Smidt P, Lecuit M, Brisse S. 2016. Whole genome-based population biology and epidemiological surveillance of *Listeria monocytogenes*. Nat Microbiol 2:16185. https://doi.org/10.1038/nmicrobiol.2016.185.
8. Mwangi MM, Wu SW, Zhou Y, Sieradzki K, de Lencastre H, Richardson P, Bruce D, Rubin E, Myers E, Siggia ED, Tomasz A. 2007. Tracking the in vivo evolution of multidrug resistance in *Staphylococcus aureus* by whole-genome sequencing. Proc Natl Acad Sci U S A 104:9451–9456. https://doi.org/10.1073/pnas.0609839104.
9. Walker TM, Merker M, Knoblauch AM, Helbling P, Schoch OD, van der Werf MJ, Kranzer K, Fiebig L, Kroger S, Haas W, Hoffmann H, Indra A, Egli A, Cirillo DM, Robert J, Rogers TR, Groenheit R, Mengshoel AT, Mathys V, Haanpera M, van Soolingen D, Niemann S, Bottger EC, Keller PM. 2018. A cluster of multidrug-resistant *Mycobacterium tuberculosis* among patients arriving in Europe from the Horn of Africa: a molecular epidemiological study. Lancet Infect Dis 18:431–440. https://doi.org/10.1016/S1473-3099(18)30004-5.
10. Carleton HA, Gerner-Smidt P. 2016. Whole-genome sequencing is taking over foodborne disease surveillance. Microbe Mag 11:311–317. https://doi.org/10.1128/microbe.11.311.1.
11. Stevens EL, Timme R, Brown EW, Allard MW, Strain E, Bunning K, Musser S. 2017. The public health impact of a publically available, environmental database of microbial genomes. Front Microbiol 8:808. https://doi.org/10.3389/fmicb.2017.00808.
12. Beerenwinkel N, Daumer M, Oette M, Korn K, Hoffmann D, Kaiser R, Lengauer T, Selbig J, Walter H. 2003. Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes. Nucleic Acids Res 31:3850–3855. https://doi.org/10.1093/nar/gkg575.
13. Votintseva AA, Bradley P, Pankhurst L, del Ojo Elias C, Loose M, Nilgiriwala K, Chatterjee A, Smith EG, Sanderson N, Walker TM, Peto TEA, Crook DW, Iqbal Z. 2017. Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples. J Clin Microbiol 55:1285–1298. https://doi.org/10.1128/JCM.02483-16.
14. Gardy JL, Loman NJ. 2018. Towards a genomics-informed, real-time, global pathogen surveillance system. Nat Rev Genet 19:9. https://doi.org/10.1038/nrg.2017.88.
15. Drummond AJ, Suchard MA, Xie D, Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 29:1969–1973. https://doi.org/10.1093/molbev/mss075.
16. To TH, Jung M, Lycett S, Gascuel O. 2016. Fast dating using least-squares criteria and algorithms. Syst Biol 65:82–97. https://doi.org/10.1093/sysbio/syv068.
17. Volz EM, Frost SDW. 2017. Scalable relaxed clock phylogenetic dating. Virus Evol 3:vex025. https://doi.org/10.1093/ve/vex025.
18. Sagulenko P, Puller V, Neher RA. 2018. TreeTime: maximum-likelihood phylodynamic analysis. Virus Evol 4:vex042. https://doi.org/10.1093/ve/vex042.
19. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. 2018. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics bty407. https://doi.org/10.1093/bioinformatics/bty407.
20. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. 2015. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 31:3691–3693. https://doi.org/10.1093/bioinformatics/btv421.
21. Ding W, Baumdicker F, Neher RA. 2018. panX: pan-genome analysis and exploration. Nucleic Acids Res 46:e5. https://doi.org/10.1093/nar/gkx977.
22. Bradley P, den Bakker H, Rocha E, McVean G, Iqbal Z. 2017. Real-time search of all bacterial and viral genomic data. bioRxiv https://www.biorxiv.org/content/early/2017/12/15/234955.
23. Argimon S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, Feil EJ, Holden MTG, Yeats CA, Grundmann H, Spratt BG, Aanensen DM. 2016. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microb Genom 2:e000093.
24. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR. 2018. Phandango: an interactive viewer for bacterial population genomics. Bioinformatics 34:292–293. https://doi.org/10.1093/bioinformatics/btx610.
25. Bielejec F, Baele G, Vrancken B, Suchard MA, Rambaut A, Lemey P. 2016. SpreaD3: interactive visualization of spatiotemporal history and trait evolutionary processes. Mol Biol Evol 33:2167–2169. https://doi.org/10.1093/molbev/msw082.
26. Libin P, Vanden Eynden E, Incardona F, Nowe A, Bezenchek A, EucoHIV Study Group, Sonnerborg A, Vandamme AM, Theys K, Baele G. 2017. PhyloGeoTool: interactively exploring large phylogenies in an epidemiological context. Bioinformatics 33:3993–3995. https://doi.org/10.1093/bioinformatics/btx535.
27. Allende C, Sohn E, Little C. 2015. TreeLink: data integration, clustering and visualization of phylogenetic trees. BMC Bioinformatics 16:414. https://doi.org/10.1186/s12859-015-0860-1.
28. Bedford T, Suchard MA, Lemey P, Dudas G, Gregory V, Hay AJ, McCauley JW, Russell CA, Smith DJ, Rambaut A. 2014. Integrating influenza antigenic dynamics with molecular evolution. Elife 3:e01914. https://doi.org/10.7554/eLife.01914.
29. Neher RA, Bedford T, Daniels RS, Russell CA, Shraiman BI. 2016. Prediction, dynamics, and visualization of antigenic phenotypes of seasonal influenza viruses. Proc Natl Acad Sci U S A 113:E1701–E1709. https://doi.org/10.1073/pnas.1525578113.
30. Crisan A, McKee G, Munzner T, Gardy JL. 2018. Evidence-based design and evaluation of a whole genome sequencing clinical report for the reference microbiology laboratory. PeerJ 6:e4218. https://doi.org/10.7717/peerj.4218.