Nextstrain automates real-time phylodynamic analysis of open data for endemic and emerging pathogens

Kimberly R. Andrews[†,1], Jennifer Chang[†,1], Cornelius Roemer[2,3], James Hadfield[1], Victor Lin[1,4], Anderson F. Brito[5], Richard Olumide Daodu[6,7], Isabel A. Joia[2], Kathryn Kistler[1,4], Allison Li[1,8], Louise H. Moncla[9], Miguel I. Paredes[1,4], Denise Kühnert[6], Laura Marcela Torres[10], Laura Voitl[2,3], Ivan Aksamentov[2,3], Emma B. Hodcroft[3,11], John Huddleston[1], John T. McCrone[1], John S. J. Anderson[1,4], Thomas R. Sibley[1], Jover Lee[1], Richard A. Neher[*,§,2,3], Trevor Bedford[*,§,1,4]

[1] Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Center, Seattle, WA, USA, [2]Biozentrum, University of Basel, Basel, Switzerland, [3]Swiss Institute of Bioinformatics, Lausanne, Switzerland, [4]Howard Hughes Medical Institute, Seattle, WA, USA, [5]Instituto Todos pela Saúde, São Paulo, SP, Brazil, [6]Center for Artificial Intelligence in Public Health Research, Robert Koch Institute, Berlin, Germany, [7]Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany, [8]Department of Biology, University of Washington, Seattle, WA, 98195 USA, [9]Department of Pathobiology, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, PA, USA, [10]Washington Department of Health, Shoreline, WA, USA, [11]Swiss Tropical and Public Health Institute, University of Basel, Allschwil, Switzerland

*Corresponding authors: richard.neher@unibas.ch and tbedford@fredhutch.org
[†]These authors contributed equally to this work.
[§]These authors contributed equally to this work.

Abstract

Motivation: Genome sequencing provides an exceptional window into the evolutionary and epidemiological dynamics of endemic and emerging pathogens, and thus allows for better, more targeted, public health interventions. Online genomic surveillance platforms can provide near real-time insight into these dynamics.

Results: Nextstrain provides continually updated real-time genomic surveillance for 21 viruses and the bacterial pathogen *Mycobacterium tuberculosis*, with most analyses relying solely on open sequence data. Each pathogen includes steps to fetch and curate open data, classify sequences using established nomenclature systems, perform phylodynamic analyses, and share the results publicly. These analyses are automated, with most running daily to provide continually updated snapshots of pathogen evolution.

Availability and Implementation: All source code is available at https://github.com/nextstrain. Phylodynamic results can be visualized and downloaded at https://nextstrain.org/pathogens, and open sequence data and curated metadata are available at https://nextstrain.org/pathogens/files.

1 Introduction

The importance of rapid, open sharing of pathogen genome sequence data has become increasingly apparent over the last decade, with major outbreaks of multiple pathogens demonstrating the value of timely access to genomic data. Open-source genomic surveillance platforms, such as Nextstrain [1], CoV-Spectrum [2] and outbreak.info [3], leverage these data to perform continually-updated analyses and make the results immediately available to the public. This process rapidly transforms shared genomic data into insights regarding pathogen dynamics, including geographic spread and emergence of new variants. This understanding enables public health agencies, epidemiologists, academics, and the broader public to respond quickly and effectively to emerging public health threats.

Nextstrain provides continually-updated and highly customizable *real-time phylodynamic* analyses for multiple pathogens of public health significance. These phylodynamic analyses integrate epidemiological data, such as sampling date and location, along with pathogen characteristics into phylogenetic analyses to provide insight into transmission dynamics [4]. We describe these as “real-time” analyses because they are kept fully up-to-date with available data, although analyses may be less up-to-date with respect to ongoing evolution, depending on the frequency at which new data becomes available. Initial Nextstrain analyses included influenza [5], dengue, Zika, and Ebola viruses. Over time, Nextstrain has expanded these analyses to include a total of 21 viral pathogens and the bacterial pathogen *Mycobacterium tuberculosis* (see Table 1). Since 2020, these real-time analyses have provided critical biological and epidemiological insights that have informed public health responses to emerging pathogens SARS-CoV-2, mpox, Oropouche, and avian influenza, while also providing ongoing analysis of endemic pathogens such as seasonal influenza and RSV.

The development and maintenance of Nextstrain’s real-time analyses require significant computational infrastructure. Here we describe the bioinformatics strategy that enables these analyses across multiple pathogens. Nextstrain’s pipelines are primarily based on *open data*, defined here as data that have been shared in a fashion that permits resharing and analyses with attribution to the original data generators. The exceptions are a subset of pipelines for SARS-CoV-2 and influenza, which are based on data from GISAID [6] and are thus restricted in re-sharing of curated data and analysis results. The open data pipelines include four main steps: 1) curation of open data using databases such as GenBank, Sequence Read Archive (SRA), and Pathoplexus; 2) quality control and clade calling; 3) customizable phylogenetic analyses; and 4) interactive visualization. These steps are automated, often running daily, and the outputs are made publicly available through nextstrain.org and via API access. The phylogenetic analysis step is highly flexible and can be tailored to run multiple different analyses for the pathogen of interest, including analyses focused on certain lineages, parts of the genome, geographic regions, or time periods. These analyses have direct surveillance utility, and can also serve as starting points for downstream analyses by public health agencies or academic labs.

2 Materials and Methods

2.1 Overview of pipeline architecture

Nextstrain real-time genomic monitoring pipelines include steps to 1) ingest and curate sequence data and associated metadata from public repositories; 2) classify sequences to clades, lineages, or genotypes; 3) perform evolutionary analyses; and 4) share and visualize results at nextstrain.org (Fig. 1). These steps rely heavily on Nextstrain software packages, including Nextclade for viral sequence classification [7], Augur for phylogenetic analyses [8], Auspice for interactive visualization, and a RESTful API for data sharing. The pipelines are automated to run at regular intervals (usually daily) using GitHub Actions to keep the analyses up-to-date as new sequences become available. Sequence data and metadata are primarily fetched from GenBank, SRA, and Pathoplexus, since these are large, centralized, open databases that are accessible through APIs. However, the pipelines can be tailored to fetch sequence data from any source that provides API access. All of our pipelines are built using Snakemake workflow manager [9], and the code for each pipeline is publicly available on GitHub, with each pathogen having a separate repository under the broader Nextstrain organization. Within this general framework, our pipelines have a number of features that differ between viral and bacterial analyses, which we describe in the following sections.

2.2 Viral analysis pipeline

For viral pathogens, our real-time analysis pipelines are divided into two Snakemake workflows called “ingest” and “phylogenetic” (Fig. 1). The *ingest workflow* fetches sequence data and metadata from external repositories, and then curates the metadata. In addition, if a Nextclade dataset is available for the pathogen of interest, the ingest workflow uses Nextclade to perform sequence quality assessment and assign lineages to each sample. The outputs of the ingest workflow are then used as inputs to the *phylogenetic workflow*, which performs subsampling, alignment, and a wide range of evolutionary analyses, including phylogenetic and phylodynamic analyses. In addition, our viral analyses include automation that runs the ingest and phylogenetic workflows on a regular basis to provide continually updated genomic monitoring. For some viral pathogens, our analysis pipelines also include a Snakemake workflow called “nextclade.” This workflow is not run as part of the continually updated real-time analyses, but instead generates a stable reference phylogeny that can be used as a Nextclade dataset. The Nextclade workflow is only updated after new lineages are designated or substantial new diversity has emerged [7] for the pathogen of interest. In the following sections we describe each of these analysis components in greater detail.

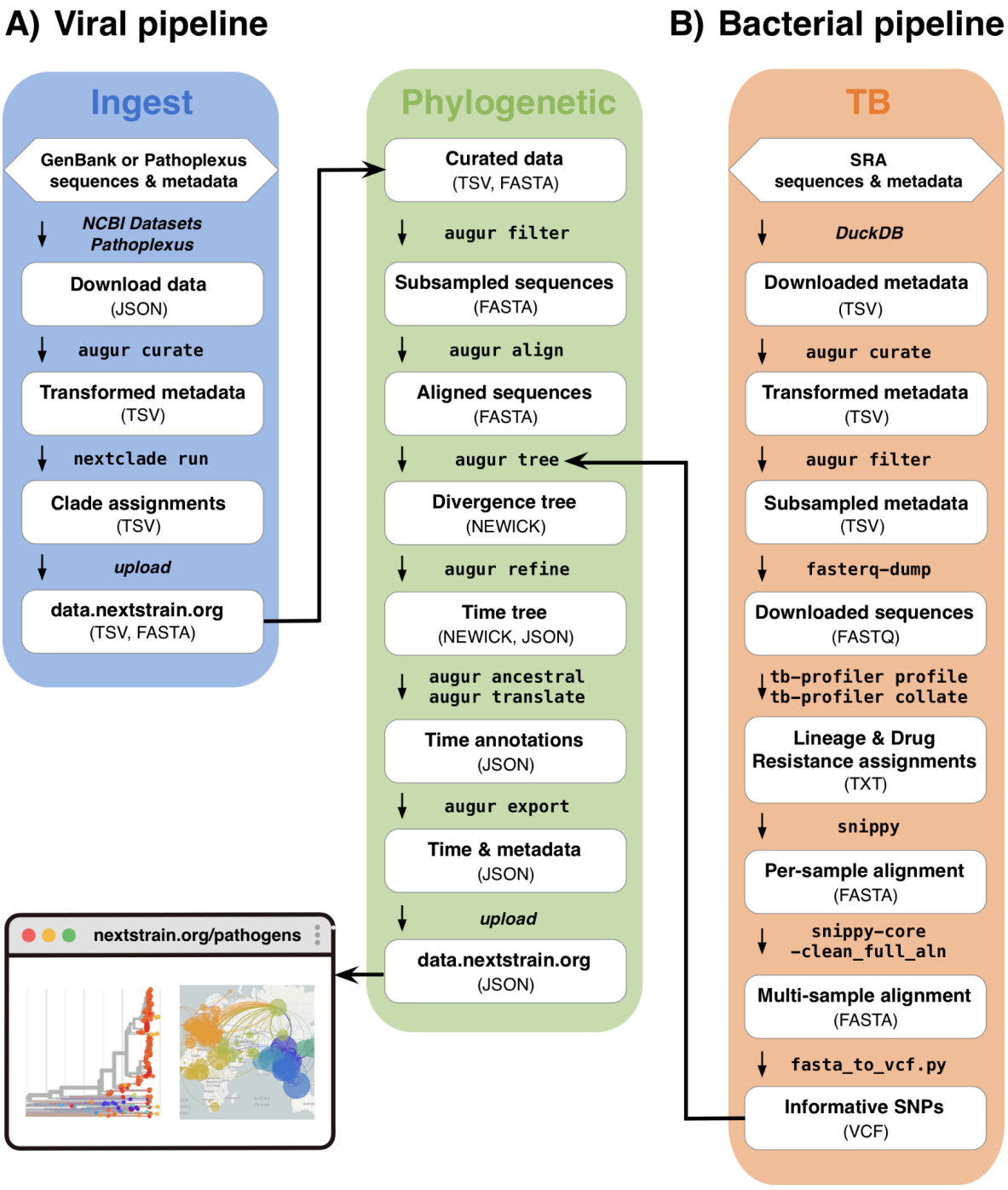

Figure 1. Outline of major steps for Nextstrain real-time analysis pipelines for viral and bacterial pathogens. Each pipeline is customized to the pathogen of interest, and the steps illustrated here represent those commonly shared across pipelines. The steps are arranged by the Snakemake workflows in which they occur within the pipelines, with each workflow shown as a colored block. The solid arrow indicates that the outputs of the ingest workflow are used as inputs for the phylogenetic workflow of the viral pipeline. The dashed arrow indicates that the final steps of the bacterial pipeline are the same as those of the phylogenetic workflow of the viral pipeline; the arrow marks the point at which the two pipelines converge. The bacterial workflow is customized for *Mycobacterium tuberculosis*.

2.2.1 Ingest workflow

The first step of the ingest workflow fetches all consensus genome sequences and associated metadata for the pathogen of interest from an external database. Most of our viral workflows currently fetch data from GenBank using the programs NCBI Datasets [10] or Entrez [11]. However, a growing number of our workflows fetch data from Pathoplexus [12]. The second step of the ingest workflow transforms the metadata to a format that is more tractable for downstream analyses and standardizes metadata values; for example, we transform collection dates to the format "YYYY-MM-DD" and parse geographic location information into standardized values that are separated into fields for country, division, and location. For viruses that have a Nextclade dataset, our ingest workflow also uses Nextclade to assign lineages and perform sequence quality assessment for each sample, and appends a column to the metadata file with the lineage assignments and quality control metrics for each sample. The final step of the ingest workflow uploads the sequences and transformed metadata to data.nextstrain.org to enable downstream phylogenetic analyses by Nextstrain workflows or external users (Fig. 3) The full list of sequence and metadata files that are publicly available for each pathogen can be viewed at nextstrain.org/pathogens/files.

2.2.2 Phylogenetic workflow

The first step of the phylogenetic workflow downloads the outputs of the ingest workflow, which include the consensus genome sequences and the harmonized metadata. The subsequent steps are highly customizable to perform analyses addressing a wide range of research questions, and most of our pipelines perform multiple analyses per pathogen. These steps primarily use Nextstrain's Augur software [8], which is a command line toolkit for pathogen phylogenetic analysis that provides built-in functionality for many data processing tasks, while also serving as a wrapper to run and standardize inputs and outputs for widely used bioinformatic tools such as MAFFT [13], IQ-TREE [14], and TreeTime [15]. Typically the first of these steps interrogates the metadata to select an appropriate subset of sequences to address the research question of interest. This subsampling step is highly customizable and includes options such as selecting or excluding sequences that match specified metadata parameters, selecting certain numbers or weighted proportions of samples across groups of metadata parameters, and setting a maximum number of total sequences. For example, a common approach is to subsample evenly over time and geography to obtain a representative worldwide subset of approximately 3000–5000 samples. After subsampling, the consensus genome sequences of the selected samples are then aligned to a reference genome, which is a representative genome sequence for the pathogen of interest; our primary criteria for selecting a reference genome are that it should be widely used and have high quality sequence data and annotations. The aligned sequences are then used to produce a maximum likelihood phylogeny with IQ-TREE and a time-resolved phylogeny with TreeTime. One of the main outputs of the phylogenetic workflow is a JSON file which contains information regarding the structure of the phylogenies, node annotations such as nucleotide and amino acid mutations, and relevant metadata for each sample. This file can be used as input to Nextstrain's Auspice software for interactive visualization of the phylogenies. The final step of the phylogenetic workflow uploads this JSON file to nextstrain.org, where it can be publicly viewed with Auspice.

2.2.3 Automation

We automate ingest and phylogenetic workflows to run at a particular cadence, generally once daily. This automation is effected through a GitHub Actions workflow that performs the following steps: 1) run the ingest workflow to download and curate data from an external repository; 2) check whether any new sequence data has been deposited in the external repository since the last run of the ingest workflow by comparing file identifiers between the previous and new download; 3) run the phylogenetic workflow only if new sequence data has been detected; and 4) post the results to nextstrain.org. These steps ensure that the computationally expensive phylogenetic analyses only run if new sequence data are available. This GitHub Action is run daily for most pathogens, but is run weekly for some pathogens to reduce computational load. These automated processes produce continually updated resources including curated data and phylogenetic datasets available for online visualization, or for download as JSON files (Fig. 2).

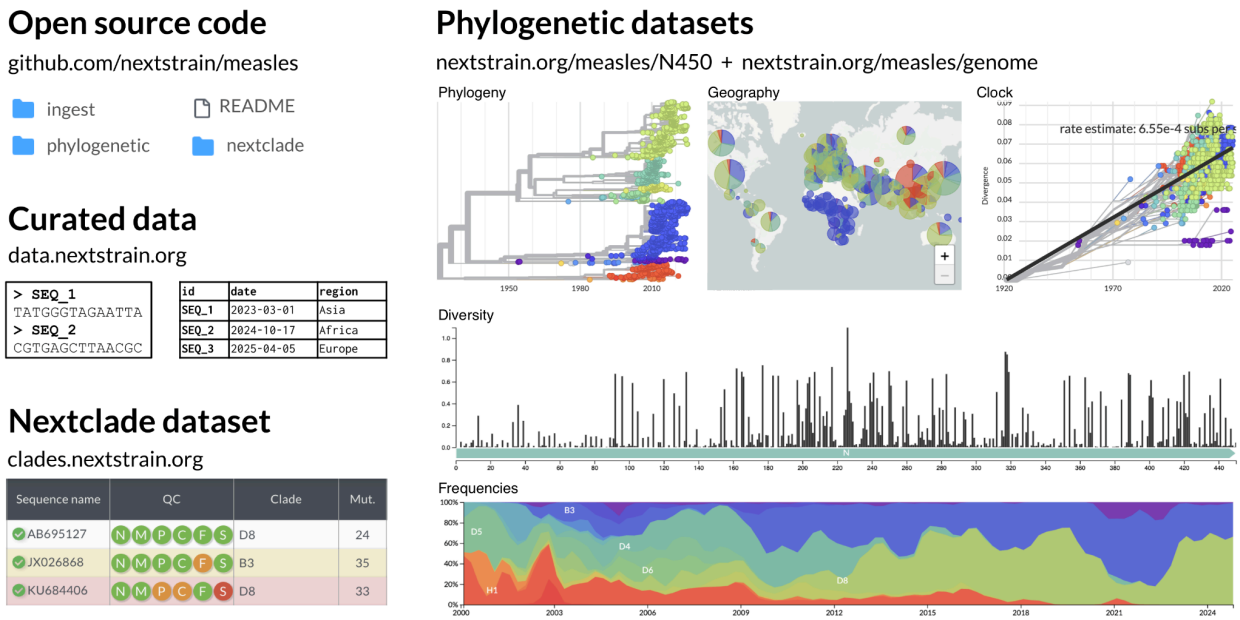

Figure 2. Example output resources for the real-time phylogenetic analysis of measles. These resources include 1) open source code for the analyses, available on GitHub; 2) sequence data and curated metadata from public repositories for the pathogen of interest; 3) a reference phylogeny that can be used as a Nextclade dataset; and 4) analytical results that can be visualized and downloaded at https://nextstrain.org.

2.2.4 Nextclade workflow

Nextclade datasets are pathogen-specific tools that align viral genomes to a reference genome, perform sequence quality assessment and clade assignment, and identify mutations relative to the reference and clade founders [7]. An integral part of a Nextclade dataset is a reference phylogeny that includes representative samples from each clade, and the *nextclade workflow* generates these reference trees (Fig. 2). The nextclade workflow is similar to the phylogenetic workflow in our real-time analyses, but has a subsampling step that focuses on obtaining a subset of complete, high quality genomes that are representative of the genetic diversity for the

pathogen of interest. Since the nextclade datasets serve as versioned and stable resources, this workflow is not run as part of the automated analyses, but can be triggered on demand.

2.2.5 Customization for virus-specific features

Although our real-time viral pathogen pipelines follow the standardized format described in the preceding sections, each pipeline is customized to accommodate the biological features and research questions relevant to each virus. For viruses with more sequence data available for certain genes or loci than for whole genomes, we produce gene- or locus-specific phylogenies as well as whole genome phylogenies (e.g., E gene for dengue, N450 for measles). For segmented viruses (e.g., Lassa, Oropouche, influenza), we build separate phylogenies for each segment of the genome rather than whole genome phylogenies, since segment reassortment strongly impacts phylogenetic inference. In addition, many of our pipelines include analyses that use customized subsampling strategies to focus on certain geographic regions, lineages, or time periods. For example, when outbreaks occur, we often add a phylogeny focused on the lineage, geographic region, and time period associated with the outbreak (e.g., mpox, avian influenza). Also, many pipelines include customized coloring options to highlight important metadata or evolutionary outcomes (e.g., host species for West Nile Virus and rabies; epitope mutations for influenza).

2.3 Bacterial analysis pipeline

The pipeline for the bacterial pathogen, *M. tuberculosis*, shares the same basic steps and outputs as the viral pipelines. However, many of these steps are implemented differently to accommodate the biological differences between bacteria and viruses, such as the much larger genome sizes for bacteria (e.g., 4.4 MB for *M. tuberculosis*) (Fig. 1). The *M. tuberculosis* pipeline is also structured to include just one Snakemake workflow to run the entire analysis, rather than including separate ingest and phylogenetic workflows, as in the viral pipelines. In addition, the automation pipeline includes a number of differences to accommodate the greater computational resources needed to run the *M. tuberculosis* analysis. In the following sections, we describe the architecture of the *M. tuberculosis* pipeline, highlighting the differences from our viral pipelines.

2.3.1 Ingest and phylogenetic workflow

One of the main differences of the *M. tuberculosis* pipeline compared to the viral pipelines is that it starts with raw short-read sequence data rather than consensus genome sequences. As a result, the *M. tuberculosis* fetches sequence data and metadata from the SRA database, whereas the viral pipelines fetch data from GenBank (Fig. 1). The pipeline starts by fetching the metadata for all *M. tuberculosis* samples with Illumina shotgun sequence data. In contrast to our viral pipelines, this step does not fetch any of the sequence data, due to the large file sizes for raw sequence data and large total number of *M. tuberculosis* samples on the SRA. Instead, the pipeline curates the metadata as described previously for the viral workflow, interrogates the transformed metadata to select a representative subset of approximately 1000 samples, and then fetches the corresponding read-level sequence data for only those samples (contained in FASTQ files).

Next, the pipeline uses Snippy (https://github.com/tseemann/snippy) to align the sequence reads to a reference genome, identify variable sites for each sample, and create a multi-sample alignment of all the genome sequences. The pipeline also implements quality control checks at two different steps to remove samples with low-quality sequence data, and performs masking of sites in the alignment that are known to be difficult to genotype in *M. tuberculosis* [16] by replacing nucleotides at those sites with ambiguous bases (N's). Next, to reduce the computational time required for downstream phylogenetic analysis, the information in the masked multi-sample alignment file is transformed into a compact VCF file that contains a summary of genotypes for all samples at phylogenetically informative sites. This file is used to build a maximum likelihood phylogeny using an Augur function that allows VCF files to be used as input to IQ-TREE.

Our *M. tuberculosis* pipeline also uses the program TBProfiler [17] to predict resistance to anti-tuberculosis drugs for each sample. This is accomplished by comparing the genome sequences of each sample against a database of mutations associated with drug resistance published by the World Health Organization and other sources [18]. TBProfiler also assigns a phylogenetic lineage to each sample using a reference database of lineage-specific mutations.

As in the viral workflow, the final output of the *M. tuberculosis* pipeline is a JSON file that can be used for interactive visualization with Auspice. The final step of the *M. tuberculosis* pipeline uploads this JSON file to nextstrain.org so that the results can be viewed publicly.

2.3.2 Automation

The *M. tuberculosis* pipeline requires substantially more computational resources than most of our viral pipelines because it stores and analyzes Illumina shotgun sequencing FASTQ files, which are much larger than the FASTA files containing consensus genome sequences that are used in our viral pipelines. Thus, our automation pipeline for the *M. tuberculosis* analysis incorporates several features to accommodate this higher computational demand. First, the pipeline uses GitHub Actions to run the workflow with AWS Batch for larger instances with more CPUs, memory, and disk space. Second, every time we run the analysis, we cache the Snippy and TBProfiler results for each sample in an AWS S3 bucket, and then if future runs select any samples that were analyzed in a previous run, we download the Snippy and TBProfiler results for those samples from the S3 bucket rather than re-running those analyses. Since each run of the workflow selects approximately 1000 samples from all *M. tuberculosis* samples in the SRA that have Illumina shotgun sequence data (currently about 170,000 samples total), the cache builds up slowly over time, resulting in decreased workflow runtime as more runs are performed. Finally, we also address the higher computational demands of the *M. tuberculosis* pipeline by running the workflow only once a week, rather than once a day as for most of our viral analyses.

2.3.3 Customization for bacteria-specific features

The *M. tuberculosis* pipeline is currently Nextstrain's only automated analysis pipeline for a bacterial pathogen, and it is highly customized for *M. tuberculosis*. For example, this workflow does not account for homologous recombination or horizontal gene transfer, which occur at very low rates in *M. tuberculosis*, but are common in many other bacterial species and can substantially influence phylogenetic inference. Potential future real-time pipelines for other bacterial species would need to include modifications to the current *M. tuberculosis* pipeline to

accommodate the unique biological features of each species, such as approaches that filter regions affected by horizontal evolution (e.g., Gubbins [19]).

2.4 Visualization of results

The results for all of our real-time analyses are publicly viewable with Auspice at nextstrain.org. Each pathogen generally includes an interactive phylogenetic tree, a map of the world with pie charts showing the geographic distribution of samples, a plot showing the diversity of genetic variation at each position along the genome, and a plot showing the frequency of different types of samples over time. These plots are highly interactive; for example, users can zoom into certain parts of the tree, change the coloring in the plots to visualize different metadata parameters, view results for a single nucleotide position, view inferred ancestral states at internal nodes such as mutations or geographic region, filter sequences based on various metadata parameters, and many other options. Different components of the visualization are tightly integrated so that, for example, zooming into a part of the phylogeny will update the other panels to reflect this subset of the data, and changing a coloring option will change the coloring in all panels. The Auspice visualization also credits the individuals who contributed the open genomic sequence data by displaying their names in several ways. First, the names of individuals who submitted sequences to the public repository can be viewed by clicking on samples at the tips of the branches of the phylogeny, and samples can also be filtered based on submitter names. In addition, a text file with submitter names for each sequence included in the phylogeny can be downloaded at the bottom of the Auspice visualization page.

2.5 Modifying pipelines for custom analyses

In addition to providing real-time genomic surveillance, our pipelines also function as starting points for external users to develop new analyses that address their own research questions. Nextstrain provides multiple tools to streamline installation and usage of our pipelines, including a command line interface (CLI) and Docker, Conda, and Singularity runtimes. In addition, our pipelines are highly customizable to address a wide variety of questions. For example, the pipelines can be modified to focus on certain phylogenetic lineages or localized geographic regions by modifying the subsampling parameters. Alternatively, users can modify the pipelines to use their own private sequence data and metadata, or a combination of private data and subsampled public data, and to enable coloring by any metadata parameter of interest. Furthermore, our pipelines can be modified for use with other pathogens (e.g., [20]). Most of our viral workflows can run on a laptop, but our *M. tuberculosis* workflow requires substantially more computational resources and typically should be run on high-performance computing (HPC) infrastructure. In addition, the *M. tuberculosis* workflow requires a different software stack than the viral workflows, and so we provide a separate Docker image for this workflow.

2.6 Sharing analysis results

Nextstrain also provides multiple methods for external users to share the results of their custom analyses publicly or privately. Users can visualize phylogenies by simply dragging and dropping JSON output files and metadata files onto the web-based tool https://auspice.us. Alternatively, Nextstrain provides two methods by which external users can visualize their results through nextstrain.org. First, any JSON output files that are stored on GitHub can automatically be viewed using nextstrain.org/community. Second, external users can set up a "Nextstrain Group" which is managed by Nextstrain and allows sharing of results at nextstrain.org/groups either publicly or privately within a group of researchers. Nextstrain also enables users to create interactive, data-driven narratives through the "Nextstrain Narrative" feature, which uses Auspice to display explanatory text alongside a corresponding view into the data that automatically updates as the user moves through the narrative [21].

3 Results

3.1 Summary of pipeline features

Nextstrain currently maintains analyses for 22 core pathogens listed in Table 1. Of these, 19 represent pathogens that have automated real-time analysis of open data, including 18 viruses and one bacteria. Two additional pipelines are not fully automated (Ebola and enterovirus D68) and one only uses restricted data (seasonal influenza). For two pathogens (avian influenza and SARS-CoV-2), a subset of analyses use restricted data (Table 1). The number of phylogenies produced by each of our pipelines ranges from one (for Zika, rabies, *M. tuberculosis*) to 87 (for SARS-CoV-2), with most pipelines producing between two and ten phylogenies (Table 1). Pipelines that generate multiple phylogenies typically do so to analyze multiple genomic regions (such as segments, genes, or loci) or to focus on multiple biologically important lineages, geographic regions, or time periods. All source code for the pipelines is available at https://github.com/nextstrain and analysis results are available at nextstrain.org/pathogens.

3.2 Examples of rapid outbreak responses

Over the last decade, Nextstrain real-time analyses have provided important insights into pathogen biology and epidemiology. The utility of these analyses has been particularly apparent during pathogen outbreaks, when rapid delivery of relevant information is critical for public health decision-making.

3.2.1 Mpox

In July of 2022, a global outbreak of mpox virus resulted in the declaration of a public health emergency of international concern (PHEIC) [22], and researchers and public health workers from around the world began sharing mpox genome sequences on GenBank. In response,

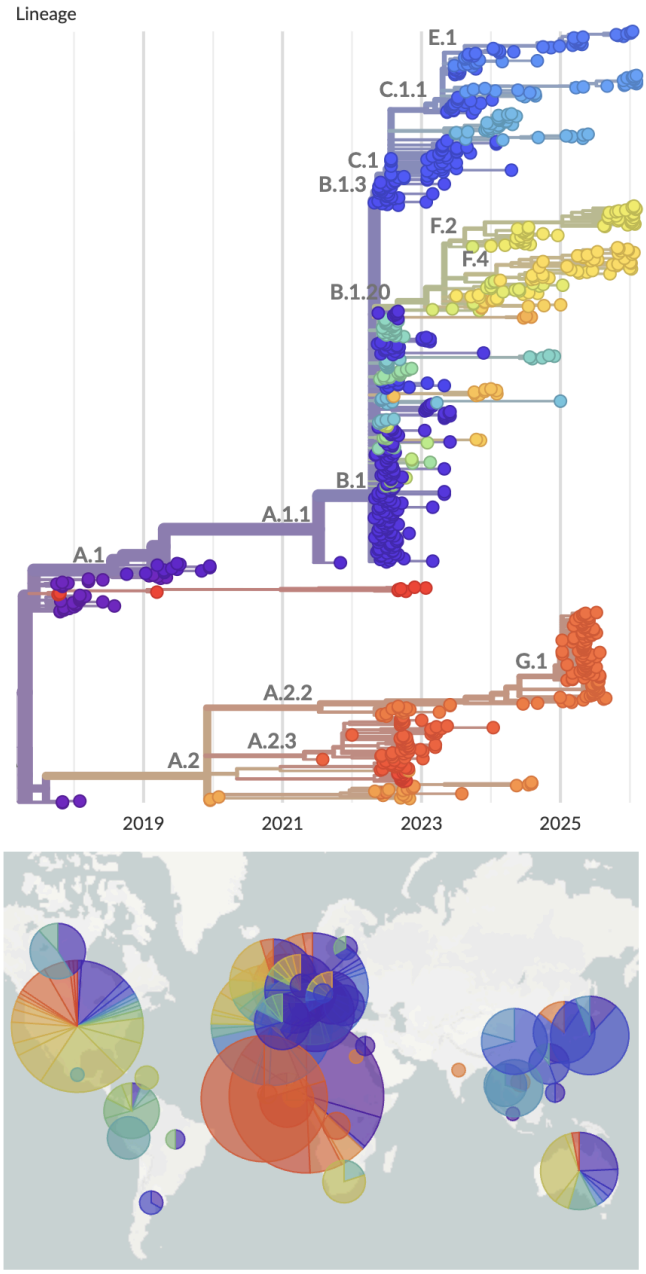

Figure 3. Phylodynamics of mpox clade IIb showing 927 viruses colored by lineage. Based primarily on APOBEC3 mutation patterns [23], this is believed to represent human-to-human transmission starting in late 2017. A live version of this analysis is viewable at nextstrain.org/mpox/clade-IIb. Screenshot is from March 9, 2026.

Nextstrain quickly developed a real-time analysis pipeline to conduct phylodynamic analyses using these sequences (Fig. 3). This pipeline required customization to accommodate several features of the mpox genome that differ from common epidemic viruses, including a much larger genome size (~200k base pairs) and the presence of large deletions and repetitive regions [23]. These features were addressed by masking repetitive genomic regions and adjusting parameter values for alignment of sequences to the reference genome. Results from this pipeline provided

timely insight into mpox transmission dynamics [24] and facilitated a nomenclature system for mpox that includes an open method for proposing new lineage designations [25], [26]. In addition, this pipeline enabled Nextstrain to quickly respond to a second PHEIC declared for mpox in 2024 [27] by developing an additional phylogenetic analysis focused on a clade that was spreading rapidly in Central Africa. These mpox resources provided situational awareness and supported downstream genomic analysis by public health and academic groups [28–30].

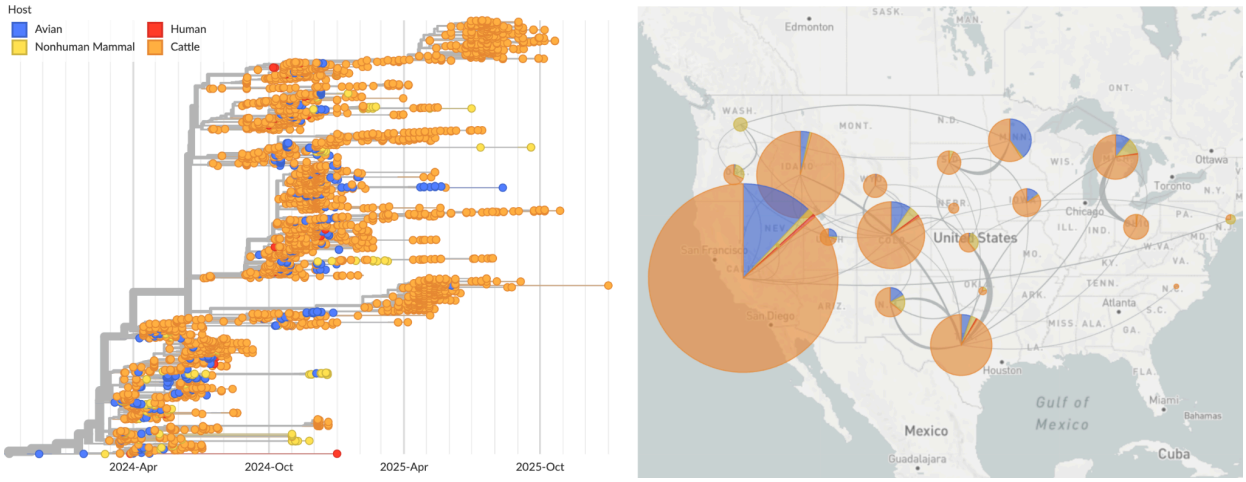

Figure 4. Full genome phylodynamic analysis of avian influenza H5N1 outbreak in North America. A live version of this analysis is viewable at nextstrain.org/avian-flu/h5n1-cattle-outbreak/genome. Screenshot is from March 9, 2026.

3.2.2 Avian influenza

When an outbreak of highly pathogenic avian influenza subtype H5N1 occurred in dairy cattle in North America in early 2024 [31], a Nextstrain real-time analysis pipeline was already in place for this pathogen, and we quickly expanded this pipeline to include new phylogenetic analyses focused on the cattle outbreak (Fig. 4). Sequence submission to public repositories was primarily by the National Veterinary Services Laboratories (NVSL) of the Animal and Plant Health Inspection Service (APHIS) of the U.S. Department of Agriculture (USDA). These submissions included both raw sequence data deposited to the SRA, and the corresponding genome assemblies deposited to GenBank, although genome assemblies were largely delayed in appearing on GenBank, with an average of 41 days between SRA and GenBank deposition. To address this issue, the Andersen Lab at Scripps Research developed an automated pipeline to assemble consensus genomes from SRA data and make the resulting genome assemblies publicly available through GitHub (https://github.com/andersen-lab/avian-influenza). In response, Nextstrain tailored its analysis pipeline to use consensus genome sequences from both GenBank and the GitHub repository. The phylodynamic analyses implemented in this pipeline provided rapid insight into the number and timing of spillover events between birds, cattle, humans, and other mammals; transmission dynamics between U.S. states; and potential evolutionary adaptations of the virus to mammalian hosts. Early in the outbreak, these analyses showed clear support for a single spillover from birds followed by transmission among cattle, highlighting a novel epidemiologic event that was distinct from outbreaks in poultry. As the outbreak in cattle grew and spread, these analyses allowed rapid identification of spillbacks of

H5N1 from cattle into domestic poultry and cats, and the ability to discern the source of infections stemming from exposure to raw milk and raw pet food.

3.3 Broader phylodynamic analyses

Beyond rapid outbreak response, analyses in Nextstrain provide an overview of the evolution and epidemiology of a number of endemic and epidemic pathogens. For example, these analyses have provided information regarding the number, timing, and geographic source of introductions of pathogens into different geographic regions. This information is important for developing effective prevention and control measures and is particularly relevant for epidemics and outbreaks of measles, mumps, Ebola, Zika and Oropouche (e.g., [32,33]). Nextstrain analyses have also been used to detect pathogen spillover to new host species, which provides important information regarding the probability of future spillover events and whether those events are likely to result in sustained transmission in the new host species (e.g., [34]). These spillover analyses are particularly relevant for rabies, avian influenza and Yellow Fever virus. Nextstrain has been used to analyze patterns of reassortment and recombination [35,36]. In addition, evolutionary analyses track the emergence and spread of new clades that have fitness advantages, and identify amino acid changes that may be responsible for those advantages [37]. These analyses can identify variants of potential biological relevance, and can improve decision making regarding strain selection for vaccine development in pathogens that undergo antigenic evolution [38].

4 Discussion

Nextstrain real-time analyses demonstrate the valuable insights that can be gained by pathogen genomic surveillance through the integration of open sequence data, open-source analytical software, and openly shared results. By combining these resources, Nextstrain now provides continually updated, publicly accessible genomic monitoring using open data for 19 important pathogens. These analyses provide up-to-date information regarding transmission dynamics, geographic spread, emergence of new variants, spillover to new host species, and other important evolutionary and epidemiological processes. This information is critical for detecting and responding to public health threats in a timely and effective manner, and these types of analyses have become indispensable parts of modern infectious disease research and outbreak response.

Nextstrain's real-time analyses also illustrate the importance of making open pathogen genomic surveillance resources easily accessible. For example, genomic sequence data are most accessible from centralized databases that provide API access, allowing any user to retrieve data programmatically; such access is critical for automating Nextstrain's real-time analysis pipelines. In contrast, databases that are password-protected or that require manual downloads through graphical interfaces need human intervention every time the analysis is updated, thereby hindering access, automation, and real-time analysis. For this reason, Nextstrain's pipelines primarily rely on GenBank, SRA, and Pathoplexus databases, since these large, open, centralized repositories provide API access.

Effective genomic surveillance also relies on accessibility of software and analysis results. Nextstrain provides API access allowing any user to download the real-time analysis outputs, including open sequences, curated metadata, and phylogenetic analysis outputs. Nextstrain also seeks to maximize accessibility by making its computational infrastructure easy for external users to install, use, and customize. In addition, Nextstrain fosters collaboration by providing multiple methods for users to share their analysis results either publicly or privately. By increasing accessibility, these features empower local public health agencies and academic laboratories to develop customized analyses that can inform responses within their own local regions. Such locally driven analyses enhance the effectiveness of public health interventions, since local communities are best positioned to understand and address local needs and constraints.

Open-source genomic surveillance platforms rely on the generous sharing of sequence data, and in turn these platforms can increase the visibility and impact of that data. With Nextstrain, we aim to ensure that data generators benefit from sharing — by increasing the reach and utility of their work while giving clear credit to the individuals behind it. GenBank, SRA, and Pathoplexus all credit individuals who submit sequence data by including submitter names in associated metadata. Pathoplexus differs from GenBank and SRA in that it includes both open and restricted sequence data; restricted data may be used in unpublished work like Nextstrain real-time analyses, but not in scientific publications or preprints. To surface this credit as broadly as possible, the curated metadata generated by our ingest workflow includes submitter names for each sequence, along with restriction status and terms of use for Pathoplexus sequences. Submitter names and restriction status for each sequence can also be viewed and downloaded directly from the Auspice visualization at nextstrain.org.

Although we focus here on Nextstrain's real-time analyses, Nextstrain is just one resource within a broader ecosystem of valuable open-source pathogen genome surveillance resources. Like Nextstrain, many of these resources are highly flexible, which often allows them to be used synergistically. For example, Nextstrain's Auspice visualization tool provides functionality for exporting and viewing phylogenetic trees in other open-source software including Taxonium [39] and MicrobeTrace [40]. Conversely, the open-source program UShER [41], which rapidly places new genome sequences on very large viral phylogenies, provides options to view smaller subtrees in Nextstrain. Similarly, Pathoplexus provides links to Nextstrain real-time analyses as a recognized resource for understanding the data housed within the repository, and other open-source software have used Nextstrain tools to aid in establishing new nomenclature systems for emerging pathogens (e.g., [25]). These examples illustrate how the development of open, flexible software can lead to interoperability across tools which further increases our ability to gain insights into pathogen dynamics.

5 Conclusion

Pathogen genomic surveillance has become a vital tool for enabling effective public health responses to emerging and endemic threats. Platforms such as Nextstrain have demonstrated the impact these tools can achieve when operating within a robust ecosystem of open data sharing among data generators and analysis groups. For such an ecosystem to function

sustainably, resources must be shared in ways that maximize accessibility while ensuring appropriate credit to those who generate them. By fostering global collaboration among public health agencies and academic laboratories, genomic surveillance can provide critical situational awareness for future outbreaks, epidemics, and pandemics.

Acknowledgements

We gratefully acknowledge the thousands of individuals who contributed to generating and sharing pathogen genomic sequences used in the real-time analyses. We also thank authors, originating laboratories, and submitting laboratories of these sequence data, as well as NCBI, Pathoplexus and GISAID for hosting databases and providing access to the data. We thank Allison Black, Vítor Borges, Sarah Cobey, Jason Caravas, Crystal Gigante, George Githinji, Peter van Heusden, Angie Hinrichs, Zamin Iqbal, Nicola Lewis, Stephanie Lunn, Placide Mbala, Laura McMullan, Marc Perry, Andrew Rambaut, Bryan Tegomoh, Sofonias Tessema, Pauline Trinh, Nídia Trovão, Mayra Trujillo, Erik Wolfson, Michael Zeller, and other public health workers and researchers for their feedback on these real-time pathogen analyses.

Funding

This work was supported by: Bill and Melinda Gates Foundation award INV-018979 to TB; NIH NIGMS R35 GM119774 to TB; US CDC contract 5 NU50CK00630; Howard Hughes Medical Institute Covid Collaboration award; John Templeton Foundation; Swiss Institute of Bioinformatics; University of Basel. TB is a Howard Hughes Medical Institute Investigator. LHM was funded with Federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services, under Contract No. 75N93021C00016.

Data availability

The sequence data and metadata used in the Nextstrain real-time analyses include all sequence data available for the 22 target pathogens in GenBank, SRA, and Pathoplexus, and all sequence data available for influenza and SARS-CoV-2 in GISAID. For data that are openly shared in these repositories, the sequence data and curated metadata are also made available through API access to data.nextstrain.org. A user interface to search these files is available at nextstrain.org/pathogens/files.

Table 1. **Summary of real-time analysis features for core pathogens.** Build endpoints = number of phylogenies produced by the pipeline; analysis size = number of samples included in the phylogenies, shown as a range when multiple phylogenies are generated; update frequency = the approximate frequency at which the automated pipeline completes the phylogenetic workflow, which depends on how often the automated pipeline is initiated and how frequently new samples are deposited in repositories. Data for this table were updated on March 5, 2026.

Core Pathogen	URL	Build Endpoints	Analysis Size	Data Provenance	Update Frequency
		Open data			
Avian influenza	avian-flu/ h5n1-cattle-outbreak/, avian-flu/h5n1-d1.1/	10	5k–6k	GenBank, SRA	Every few days
Dengue	dengue/	11	1.5k–4k	GenBank	Every few days
Ebola	ebola/	2	400–1.5k	GenBank, Pathoplexus	Monthly
Enterovirus	enterovirus/	2	1.6k–3.5k	GenBank	Rarely
Lassa	lassa/	3	700–1100	GenBank	Weekly
Measles	measles/	2	2–3k	GenBank, Pathoplexus	Every few days
Metapneumovirus	hmpv/	9	800–3k	GenBank, Pathoplexus	Weekly
Mpox	mpox/	4	1k–5k	GenBank, Pathoplexus	Every few days
Mumps	mumps/	3	500–900	GenBank	Weekly
Nipah	nipah/	4	20–100	GenBank	Weekly
Norovirus	norovirus/	14	200–6k	GenBank	Every few days
Oropouche	oropouche/	3	750	GenBank	Monthly
Rabies	rabies/	1	3k	GenBank	Weekly
RSV	rsv/	18	2.5k–3.2k	GenBank, Pathoplexus	Weekly

Rubella	rubella/	2	150–3k	GenBank	Weekly
SARS-Cov-2	ncov/open/	29	250–4.5k	GenBank, RKI	Weekly
Seasonal CoV	seasonal-cov/	4	170–750	GenBank	Weekly
Tuberculosis	tb/	1	1100	SRA	Weekly
West Nile Virus	WNV/	6	1–3k	GenBank	Every few days
Yellow Fever Virus	yellow-fever/	2	350–900	GenBank	Weekly
Zika	zika/	1	1k	GenBank	Weekly
		Closed data			
Avian influenza	avian-flu/	48	1.4–3.4k	GISAID	Monthly
SARS-CoV-2	ncov/gisaid/	58	200–4.5k	GISAID	No longer updated
Seasonal influenza	seasonal-flu/	42	250–3k	GISAID	Weekly

References

1. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. Bioinformatics. 2018;34: 4121–4123.

2. Chen C, Nadeau S, Yared M, Voinov P, Xie N, Roemer C, et al. CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants. Bioinformatics. 2022;38: 1735–1737.

3. Gangavarapu K, Latif AA, Mullen JL, Alkuzweny M, Hufbauer E, Tsueng G, et al. Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. Nat Methods. 2023;20: 512–522.

4. Grenfell BT, Pybus OG, Gog JR, Wood JLN, Daly JM, Mumford JA, et al. Unifying the epidemiological and evolutionary dynamics of pathogens. Science. 2004;303: 327–332.

5. Neher RA, Bedford T. Nextflu: Real-time tracking of seasonal influenza virus evolution in humans. Bioinformatics. 2015;31: 3546–3548.

6. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill. 2017;22.

7. Aksamentov I, Roemer C, Hodcroft E, Neher R. Nextclade: clade assignment, mutation calling and quality control for viral genomes. J Open Source Softw. 2021;6: 3773.

8. Huddleston J, Hadfield J, Sibley TR, Lee J, Fay K, Ilcisin M, et al. Augur: a bioinformatics toolkit for phylogenetic analyses of human pathogens. J Open Source Softw. 2021;6: 2906.

9. Köster J, Rahmann S. Snakemake-a scalable bioinformatics workflow engine. Bioinformatics. 2018;34: 3600.

10. O'Leary NA, Cox E, Holmes JB, Anderson WR, Falk R, Hem V, et al. Exploring and retrieving sequence and metadata for species across the tree of life with NCBI Datasets. Sci Data. 2024;11: 732.

11. Sayers EW, Bolton EE, Fine AM, Kelly C, Kim S, Landrum M, et al. Database resources of the National Center for Biotechnology Information in 2026. Nucleic Acids Res. 2025.

12. Dalla Vecchia E. Pathoplexus: towards fair and transparent sequence sharing. Lancet Microbe. 2024;5: 100995.

13. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30: 772–780.

14. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32: 268–274.

15. Sagulenko P, Puller V, Neher RA. TreeTime: Maximum-likelihood phylodynamic analysis. Virus Evol. 2018;4: vex042.

16. Marin M, Vargas R, Harris M, Jeffrey B, Epperson LE, Durbin D, et al. Benchmarking the empirical accuracy of short-read sequencing across the M. tuberculosis genome. Bioinformatics. 2022;38: 1781–1787.

17. Phelan JE, O'Sullivan DM, Machado D, Ramos J, Oppong YEA, Campino S, et al. Integrating informatics tools and portable sequencing technology for rapid detection of resistance to anti-tuberculous drugs. Genome Med. 2019;11: 41.

18. Verboven L, Phelan J, Heupink TH, Van Rie A. TBProfiler for automated calling of the association with drug resistance of variants in Mycobacterium tuberculosis. PLoS One. 2022;17: e0279644.

19. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015;43: e15.

20. Eaton K, Featherstone L, Duchene S, Carmichael AG, Varlık N, Golding GB, et al. Plagued by a cryptic clock: insight and issues from the global phylogeny of Yersinia pestis. Commun Biol. 2023;6: 23.

21. Hadfield J, Brito AF, Swetnam DM, Vogels CBF, Tokarz RE, Andersen KG, et al. Twenty

years of West Nile virus spread and evolution in the Americas visualized by Nextstrain. PLoS Pathog. 2019;15: e1008042.

22. World Health Organization. Second meeting of the International Health Regulations (2005) (IHR) Emergency Committee regarding the multi-country outbreak of monkeypox. 23 Jul 2022. Available: https://www.who.int/news/item/23-07-2022-second-meeting-of-the-international-health-regulations-(2005)-(ihr)-emergency-committee-regarding-the-multi-country-outbreak-of-monkeypox

23. O'Toole Á, Neher RA, Ndodo N, Borges V, Gannon B, Gomes JP, et al. APOBEC3 deaminase editing in mpox virus as evidence for sustained human transmission since at least 2016. Science. 2023;382: 595–600.

24. Paredes MI, Ahmed N, Figgins M, Colizza V, Lemey P, McCrone JT, et al. Underdetected dispersal and extensive local transmission drove the 2022 mpox epidemic. Cell. 2024;187: 1374–1386.e13.

25. Happi C, Adetifa I, Mbala P, Njouom R, Nakoune E, Happi A, et al. Urgent need for a non-discriminatory and non-stigmatizing nomenclature for monkeypox virus. PLoS Biol. 2022;20: e3001769.

26. Ruis C, Lusamaki E, O'Toole A, Otieno JR, Colquhoun R, Roemer C, et al. A systematic nomenclature for mpox viruses causing outbreaks with sustained human-to-human transmission. Nat Med. 2025;31: 2854–2858.

27. World Health Organization. WHO Director-General declares mpox outbreak a public health emergency of international concern. 14 Aug 2024. Available: https://www.who.int/news/item/14-08-2024-who-director-general-declares-mpox-outbreak-a-public-health-emergency-of-international-concern

28. Borges V, Duque MP, Martins JV, Vasconcelos P, Ferreira R, Sobral D, et al. Viral genetic clustering and transmission dynamics of the 2022 mpox outbreak in Portugal. Nat Med. 2023;29: 2509–2517.

29. Lau KM, Banks M, Bryant K, Lambert JD, Torres LM, Lunn SM, et al. Real-Time Use of Monkeypox Virus Genomic Surveillance, King County, Washington, USA, 2022-2024. Emerg Infect Dis. 2025;31: 76–79.

30. Akther S, Su M, Wang JC, Amin H, Taki F, De La Cruz N, et al. Genomic epidemiology of mpox virus during the 2022 outbreak in New York City. Nat Commun. 2025;16: 8354.

31. Nguyen T-Q, Hutter CR, Markin A, Thomas M, Lantz K, Killian ML, et al. Emergence and interstate spread of highly pathogenic avian influenza A(H5N1) in dairy cattle in the United States. Science. 2025;388: eadq0900.

32. Black A, Moncla LH, Laiton-Donato K, Potter B, Pardo L, Rico A, et al. Genomic epidemiology supports multiple introductions and cryptic transmission of Zika virus in Colombia. BMC Infect Dis. 2019;19: 963.

33. Moncla LH, Black A, DeBolt C, Lang M, Graff NR, Pérez-Osorio AC, et al. Repeated introductions and intensive community transmission fueled a mumps virus outbreak in

Washington State. eLife. 2021;10.

34. Damodaran L, Jaeger AS, Moncla LH. Ecology and spread of the North American H5N1 epizootic. Nature. 2026;649: 432–441.

35. Daodu RO, Chang J, Prescott J, Reinert K, Kühnert D. Lassa virus live tracking and lineage assignment: how Nextstrain can enhance surveillance and public health in Africa and beyond. Emerg Microbes Infect. 2026; 2640699.

36. Potter BI, Kondor R, Hadfield J, Huddleston J, Barnes J, Rowe T, et al. Evolution and rapid spread of a reassortant A(H3N2) virus that predominated the 2017–2018 influenza season. Virus Evolution. 2019;5.

37. Kistler KE, Huddleston J, Bedford T. Rapid and parallel adaptive mutations in spike S1 drive clade success in SARS-CoV-2. Cell Host Microbe. 2022;30: 545–555.e4.

38. Morris DH, Gostic KM, Pompei S, Bedford T, Łuksza M, Neher RA, et al. Predictive modeling of influenza shows the promise of applied evolutionary biology. Trends in Microbiology. 2018;26: 102–118.

39. Sanderson T. Taxonium, a web-based tool for exploring large phylogenetic trees. Elife. 2022;11. doi:10.7554/eLife.82392

40. Campbell EM, Boyles A, Shankar A, Kim J, Knyazev S, Cintron R, et al. MicrobeTrace: Retooling molecular epidemiology for rapid public health response. PLoS Comput Biol. 2021;17: e1009300.

41. Turakhia Y, Thornlow B, Hinrichs AS, De Maio N, Gozashti L, Lanfear R, et al. Ultrafast Sample placement on Existing tRees (UShER) enables real-time phylogenetics for the SARS-CoV-2 pandemic. Nat Genet. 2021;53: 809–816.