PERSPECTIVE
https://doi.org/10.1038/s41591-020-0935-z

nature medicine

Ten recommendations for supporting open pathogen genomic analysis in public health

Allison Black[1,2], Duncan R. MacCannell[3] ✉, Thomas R. Sibley[2] and Trevor Bedford[1,2]

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

Increasingly, public-health officials are using pathogen genomic sequence data to support surveillance, outbreak response, pathogen detection and diagnostics[1]. Sequencing cuts across traditional pathogen boundaries; for example, it can be used to distinguish cases of 'wild' polio from vaccine-derived polio[2], to predict the susceptibility of a tuberculosis infection to antibiotics[3] or to trace the source of a foodborne infection[1]. Most recently, scientists and public-health agencies are using sequence data of the coronavirus SARS-CoV-2 to investigate the origin of this virus[4,5], the global expansion of the epidemic[6] and community transmission of COVID-19 in various localities[7–10].

Because of its utility, public-health agencies throughout the world are developing their capacity to perform genomic sequencing. Almost every infectious disease program within the US Centers for Disease Control and Prevention generates and analyzes pathogen sequence data[11]. Many international public-health agencies, such as Public Health England, the Public Health Agency of Canada and the European Centre for Disease Prevention and Control, also have large sequencing programs. Capacity for pathogen genome sequencing has also grown within agencies at the state and local level. Indeed, every state public-health lab in the USA, as well as public-health labs in most major counties, conduct pathogen genome sequencing for foodborne surveillance, if not for other diseases as well.

Although laboratory capacity to generate sequence data has increased greatly, the capacity to assemble, analyze and interpret genomic data has been harder to develop. Possibly this is because many of the tools developed for sequence assembly and analysis either are expensive to license or require a high level of computational proficiency to use. Individuals with specialized training in bioinformatics and genomic epidemiology are relatively new to applied public health, and this workforce has not been distributed evenly across agencies. To grow analytic capacity, we believe that researchers must work from both ends: make bioinformatics and genomic analysis more accessible to non-specialists, and also build a larger workforce with experience in bioinformatics and genomic epidemiology within public health.

In this Perspective, we make recommendations for building a sustainable informatic infrastructure for pathogen genomics that can be used across public-health programs. We have centered our recommendations around what we feel are the fundamental characteristics of an open ecosystem for pathogen genomic analysis: reproducibility, such that genomic analysis is standardized and repeatable across agencies and through time; accessibility, at varying levels of both economic resources and technical knowledge; flexibility, providing a set of modular tools to analyze, explore and visualize genomic data across a range of public-health applications; and auditability, ensuring that genomic assembly and analysis can be validated according to strict public-health standards.

Our recommendations are not a checklist—rather, we aim to provide a structured view of what a public-health informatics ecosystem might look like (Fig. 1). We hope this work provides a starting point for the community to use to come together in designing and developing this ecosystem.

Methodology for achieving a consensus

To investigate the current landscape of bioinformatics and genomic epidemiology in public-health agencies, we conducted a series of long-form, semi-structured interviews with bioinformaticians, laboratory microbiologists performing sequencing, software engineers developing pipelines and workflow-management software for public health, and epidemiologists acting upon inferences from genomic data. We aimed to get a broad perspective, interviewing individuals from different countries, working on a wide array of pathogens and working in agencies with varied capacity for performing genomic analysis. A full list of sources of interviewees is in Table 1.

The interviews focused on the following topics: technical components of genomic analysis, considerations for genomic analysis specific to public-health settings and social issues surrounding

[1]Department of Epidemiology, University of Washington, Seattle, Washington, USA. [2]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. [3]Office of Advanced Molecular Detection, Centers for Disease Control and Prevention, Atlanta, Georgia, USA. ✉e-mail: fms2@cdc.gov

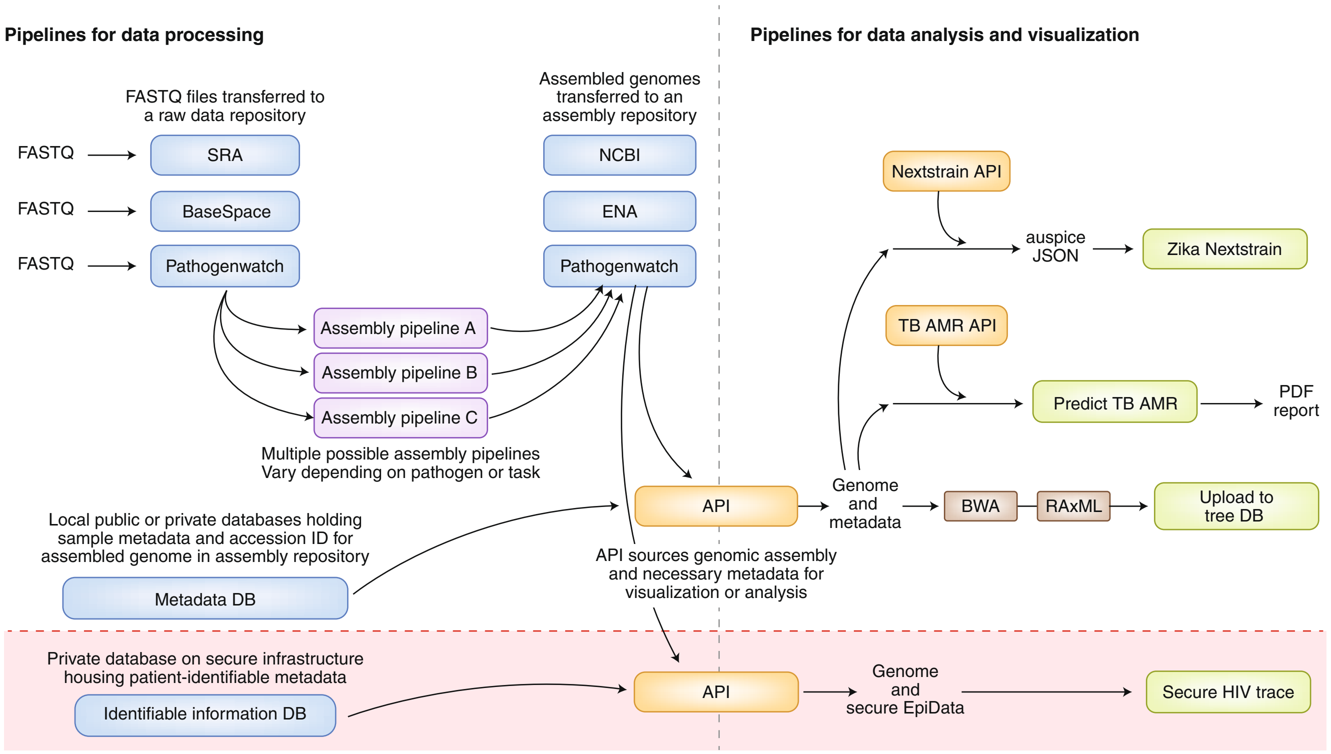
Pipelines for data processing
Pipelines for data analysis and visualization
FASTQ files transferred to a raw data repository
Assembled genomes transferred to an assembly repository
FASTQ
FASTQ
FASTQ
SRA
BaseSpace
Pathogenwatch
NCBI
ENA
Pathogenwatch
Assembly pipeline A
Assembly pipeline B
Assembly pipeline C
Multiple possible assembly pipelines
Vary depending on pathogen or task
Local public or private databases holding sample metadata and accession ID for assembled genome in assembly repository
Metadata DB
API
API sources genomic assembly and necessary metadata for visualization or analysis
Nextstrain API
auspice JSON
Zika Nextstrain
TB AMR API
Predict TB AMR
PDF report
Genome and metadata
BWA
RAxML
Upload to tree DB
Private database on secure infrastructure housing patient-identifiable metadata
Identifiable information DB
API
Genome and secure EpiData
Secure HIV trace

Fig. 1 | Data processing, analysis and visualization. Left, data processing: First, bioinformaticians must process raw reads and assemble genomes. Within the data-processing side of this ecosystem, we envisage three types of databases: one for archiving or holding raw sequencing reads, one for archiving assembled data and one for holding metadata about the samples. Various current databases could fill these positions, or new databases could be developed if public-health programs require additional utility. We imagine that the Sequence Read Archive would continue to serve as the primary raw reads database. But, for instance, for a metagenomic sample containing both pathogen and human reads, the reads could be held in Illumina's BaseSpace platform instead. From here, bioinformaticians could assemble genomes using open-access pipelines available from a cloud-based deployment platform; pipeline choice would be based around what type of assembly the user needed. Within each assembly pipeline, the final step should be automatic submission of the genome to the relevant database for that assembly type. This database could be an NCBI database (e.g., NCBI Nucleotide, NCBI Pathogen Detection), a member of the International Nucleotide Sequence Database Collaboration (e.g., DDBJ, ENA) or a pathogen-specific assembly database (e.g., GISAID, ViPR). Alternatively, if the genome sequence itself represents highly sensitive data, this database could also be a private repository available only to individuals within the public-health institution and vetted partners. The sequence data accession identifier should be deposited into a third database, the metadata database. In our design, the metadata database would be an in-house relational database that facilitates sample tracking and houses all relevant clinical and laboratory data according to a well-defined schema that can also accommodate long-form entries. Likely, it would be easier to licence databasing software for the metadata database than to build it from scratch. Importantly, metadata databases could also be secured, and could house relevant PII collected during epidemiologic investigations. Keeping these data separate from the genomic data will ensure that PII can be kept private when necessary. Data linking would occur via API calls: calls to the metadata database would pull relevant sample information and the assembly accession number, which an API would then use to source the genome assembly from the genomic database. Various metadata and genomic data combinations could be sourced depending on what data fields were necessary for the desired analytic or visualization pipeline. Right, data analysis and visualization: Once genomic assemblies and relevant metadata were combined, they could be piped to various analytic workflows, for example for predicting antimicrobial resistance, making specific data structures such as phylogenetic trees, or preparing datasets or data objects to serve as interactive data visualization platforms. We imagine that a wide array of different visualization and analytic pipelines will be in use; good APIs and complete, standardized metadata are necessary to support that breadth. Some analytic pipelines may be completely containerized, end-to-end workflows that produce visualizations or reports. Others could make data objects, such as phylogenetic trees, and submit them to an additional database for use in subsequent analyses. Additionally, analytic pipelines could make API calls to external databases, such as antimicrobial resistance gene databases, facilitating the integration of these new pipelines with existing software packages and databases.

genomic data. The interviews revealed various themes and consistent challenges related to supporting pathogen genomic analysis in public-health agencies. Our recommendations seek to address those challenges, and describe strategies for building a sustainable, efficient and effective bioinformatics infrastructure for the growing need in public health.

Although we conducted interviews primarily with the staff of public-health programs within the USA, our colleagues at the Africa Centres for Disease Control and Prevention led a concurrent effort to assess sequencing and bioinformatics capacity within African public-health agencies. We reviewed each others' landscape analyses, finding many challenges within small public-health institutions in the USA that were similar to those that exist in Africa. To ensure that our recommendations would be relevant across income settings, our colleagues at the Africa Centres for Disease Control and Prevention reviewed the recommendations outlined here for their appropriateness to public-health settings in low- and middle-income countries.

Recommendations

In this section we describe our ten recommendations for supporting open pathogen genomic analysis in public-health settings.

Table 1 | Agencies, programs and development teams participating in long-form interviews in the generation of this consensus statement

Category	Agency or team
US Centers for Disease Control (CDC)—bacterial pathogens	NCEZID/DHCPP/Bacterial Special Pathogens Branch
	NCIRD/Division of Bacterial Diseases
	NCEZID/Division of Healthcare Quality Promotion
	NCHHSTP/Division of STD Prevention–Gonorrhea and Chlamydia
	NCEZID/Division of Foodborne, Waterborne, and Environmental Diseases
	NCHHSTP/Division of Tuberculosis Elimination
CDC—viral pathogens	NCEZID/DHCPP/Viral Special Pathogens
	NCIRD/Division of Viral Diseases
	NCIRD/Influenza Division
	NCHHSTP/Division of Viral Hepatitis Prevention
	NCHHSTP/Division of HIV/AIDS Prevention
CDC—parasitic pathogens and mycoses	CGH/Division of Parasitic Diseases and Malaria
	NCEZID/DFWED/Mycotic Diseases Branch
US state public-health laboratories	Colorado Department of Public Health and Environment
	Minnesota Department of Health Public Health Laboratory
	Utah Unified State Laboratories Public Health
	Virginia Division of Consolidated Laboratory Services
	Washington State Public Health Laboratories
International public-health agencies	European Centre for Disease Prevention and Control
	Public Health Agency of Canada
Software project teams	BioNumerics
	IRIDA
	INNUENDO

Support data hygiene and interoperability by developing and adopting a consistent data model. Pathogen isolates need context. Who was the sample collected from? When was it collected? How was it collected? Without this information, often referred to as metadata, much of the value of the sample is lost, both from a clinical reporting and a data analysis standpoint. Despite the value of metadata, in current practice sequences are frequently decoupled from the full constellation of epidemiologic data and sample data that describe them (Box 1).

To ameliorate this problem, widely used genomic databases such as the Sequence Read Archive have standards and formatting requirements for submissions. However, we continue to face challenges of data incompleteness and lack of consistency in data reporting. Data incompleteness compromises the analytic utility of the data, and data inconsistency impacts users' ability to interact with the data through computer programs. As the increasing amounts of data reduce our ability to manually interact with those data, both complete and structured data will be fundamental to an informatic ecosystem that works effectively and efficiently at scale.

To improve this situation, we recommend adopting a data model (Box 1) that specifies necessary data elements and provides an appropriate structure for linking sequence data, clinical data and epidemiologic information. The required data elements specified by the data model should be sufficiently flexible that they are applicable across a wide array of pathogens. To structure the data recorded within the data model, public-health programs will also need to adopt and/or develop ontologies: controlled vocabularies that standardize free-form epidemiologic information about cases and their exposures. Standardizing how data are recorded facilitates programmatic interaction with databases, enabling users to automate quality control and analytic procedures. Two good examples of epidemiologic ontologies are the

Box 1 | The need for data context and structure

Many sequences can map to a single set of epidemiologic data in a way that is complex and often hierarchical. We illustrate that complexity here by imagining some of the genomic sequences that could be collected from a single individual infected with influenza.

When a clinician draws a single diagnostic specimen from an influenza-infected patient, that single sample may yield many distinct pieces of genomic data. For instance, we can ascertain the consensus genome of the infecting strain by sequencing the clinical isolate. From that same sequencing run there may also be separate SNP calls that describe within-host minority variants. Additionally, a lab could decide to culture the infecting strain and sequence the cultured isolates after different numbers of passages. Each of these scenarios yield sequences that are distinct from one another, and have different laboratory-associated data, but share the same epidemiologic data.

Patient data, such as their demographic information, clinical data about their illness, and exposure information describing how they may have contracted the disease, form an unchanging set of characteristics describing that individual at that point in time. It is critically important to keep this set of data, frequently referred to as metadata, linked to the genomic sequence(s) from that individual's specimen. Access to those linked data can help us explore important laboratory, clinical and epidemiological questions, such as: how did the virus change while being passaged in tissue culture? How did the strain of influenza change within a single infected individual over time? Or, what does the changing frequency of major and minor viral variants tell us about transmission? But keeping those data linked is challenging because of the complex relationships between them.

We recommend using a hierarchical model to link all of the related pieces of data that describe the patient and their infection. By explicitly describing the hierarchical relationships between constituent pieces of data, this type of data model provides a clear structure to keep data linked. In our example data model, there are three major data fields: case, sample and sequence. Each case record can have multiple samples, and each sample can have multiple sequences (Fig. 2). Linkage between the fields is maintained by a case identifier and sample identifier that are logged as subfields. Within each field, subfields record the information most pertinent to that field. For example, the case field contains the following subfields: host species, age, sex, symptoms and geographic information. The sample field contains the case identifier as a subfield, and also logs information such as sample collection date, collection medium (blood, urine, tissue) and culture information. Finally, each sample may have multiple sequences. The sequence field includes the case identifier and the sample identifier as subfields, but again organizes information more pertinent to the sequences themselves, such as what portion of the genome the sequence is from, whether the sequence is a consensus sequence or a minor variant, flags specifying whether the sequence is public or private, and the accession number for accessing the sequence within a genomic assembly database.

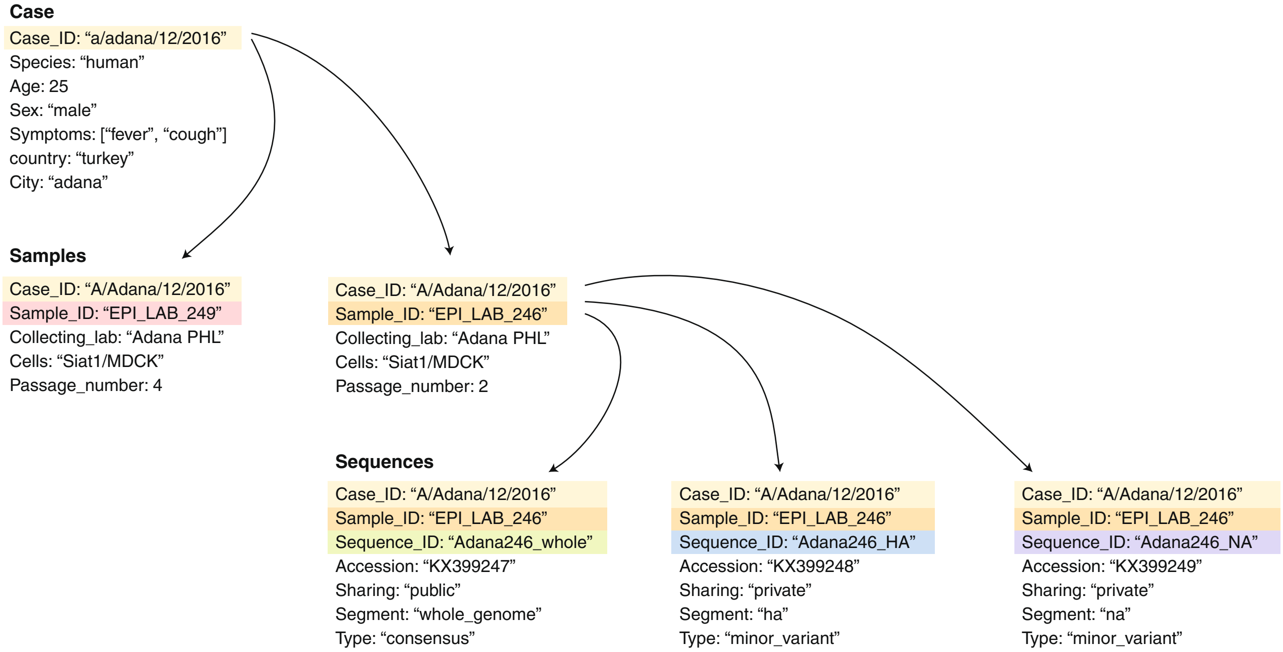

Fig. 2 | Schematic illustrating an example data model with the three major data fields: case, sample and sequence. We also show potentially relevant subfields within each field.

Integrated Rapid Infectious Disease Analysis genomic epidemiology ontology, GenEpiO (https://genepio.org/) and FoodOn[12].

Strengthen application programming interfaces. Application programming interfaces, or APIs, are the mechanism by which users communicate with computers, code and databases in an automated way. They are critically important for programmatic querying of databases, collation of disparate data sources and communication between pieces of software within a greater ecosystem.

The relative paucity of consistent and well-documented APIs for software tools and databases used by public-health bioinformaticians has at least two effects. First, the lack of APIs limits the scalability of bioinformatics analyses. Currently, querying genomic databases frequently requires human interaction via a web-based graphical user interface. However, with ever-increasing amounts of data, the ability to manually explore, source and distribute data will decline. Bioinformaticians at public-health agencies will need to automate querying and analysis; the quality of APIs will directly affect their ability to do this reproducibly and efficiently. Second, the lack of APIs leads to inefficient use of bioinformatician effort. When basic pipelines do not run automatically, or linking programs together requires considerable effort, bioinformaticians spend large amounts of time writing interstitial code and managing file format conversions. This takes up time that bioinformaticians and genomic epidemiologists could otherwise spend analyzing the data, probably with greater public-health impact.

The development and use of well-documented APIs will underlie the success of a software ecosystem within public health, and cannot be an afterthought. We recommend that public-health institutions adopt common API standards and carry out API development in tandem with database or software development. For the many software programs and databases that already exist, specific funding sources should be allocated to build or extend current APIs to function with the agreed-upon data models and adhere to adopted API standards (Box 2). Notably, the US General Services Administration has developed standards for APIs, which provide a concrete starting point in the development of APIs for genomic and epidemiologic databases. These standards are described in detail at https://github.com/GSA/api-standards.

Box 2 | Technical recommendations for APIs

Although a full discussion of API architectures is beyond the scope of this paper, in line with General Services Administration guidelines we recommend that APIs be RESTful[15]. We also recommend that APIs have clear, human-readable endpoints. They should return JavaScript Object Notation (JSON) objects for both API responses and error messages, as JSON is the standard, widely-supported structure for transferring data between programs. APIs should be versioned, and they should be backward-compatible within a major version such that updates to the API do not frequently break software and pipelines. All APIs should use HTTPS, which improves data security during transfer between client and server. Finally, APIs should have clear, readable documentation, and they should be developed transparently in an environment that allows users to ask questions, give feedback and report issues.

Develop guidelines for management and stewardship of genomic data. The increasing abundance of longitudinally collected pathogen genomic sequence data is a valuable resource for public health. To fully realize the value of this data, however, programs will need to manage and care for the data in a unified manner. To this end, public-health institutions should develop and adopt guidelines and standards for data collection, annotation, archiving, and reuse. The community should design these guidelines to ensure that data adhere to FAIR principles: that is, they are findable, accessible, interoperable and reusable[13]. Following these principles ensures that once generated, data can be reused in the future.

We recommend that guidelines describe which data to archive, including both raw and assembled data such as consensus genome sequences; the duration of archiving; systems for long-term archiving;

Box 3 | Deploying, supporting, and governing an open bioinformatics ecosystem

There are various potential ways to handle open pipeline deployment. We envisage a model with three components: a registry that catalogues available pipelines and datasets, a centralized location or 'hub' where packages and pipelines are hosted, and a graphical user interface that provides an easy-to-use portal to the hub and its pipelines.

The Registry: The registry is a directory that tells users which pipelines are available and gives information about each one. A registry entry should describe what process the pipeline performs, what inputs it takes and what outputs it provides, as well as information about which agency hosts the pipeline and how to access it.

The Package and Pipeline Hub: Pipelines and software packages could be containerized and hosted on a container hub (e.g., Docker or Singularity). This would allow individuals familiar with command-line interfaces to source and run a pipeline from the container hub using minimal shell scripts or pull and run commands.

The Portal: An open-source initiative could be funded to write and maintain graphical user interfaces for interacting with the hub or the pipelines. These interfaces, which would make it easier for non-bioinformaticians to select and run a pipeline, could either wrap the process of sourcing a pipeline from the hub and running it, or wrap pipelines that are hosted directly on a specific server. These options are not mutually exclusive, and we imagine that both a broad registry and a more narrow shared computational service that hosts the most frequently used pipelines would coexist.

The genomics community could deploy instances of the platform on distinct computing infrastructures managed by various public-health agencies or networks of agencies. The computing infrastructure could be cloud based or could use an in-house cluster. An in-house computational workforce could create and manage instances of the platform, or this work could be performed as Software as a Service, whereby a nonprofit or a company charges individual users based on their use of the platform. Notably, open-source platforms with SaaS options have been successful, as in the case of Arvados (https://arvados.org).

We emphasize that the deployment platform should support multiple generations of individual pipelines, clearly indicating which versions are vetted reference pipelines, which pipelines are under development for future release and which pipelines have been deprecated. This versioning will allow bioinformaticians and software developers to improve and develop future pipelines while maintaining access to reference pipelines, and will help users know which pipelines they should use.

To realize this system, we will need to decide which agencies should develop and host the deployment platform, containerized pipelines and access portal. We will also need to extend current pipelines and software packages such that they work in the deployment platform environment. Governance duties will include communicating the requirements of the community, such as data models, API standards and documentation standards, to developers and users.

This will take effort. In our experience, many open-source projects and software are critically important, yet are underfunded and their developers overtaxed. We emphasize that building the type of ecosystem we have proposed will require large funding sources to support initial development and sustain ongoing maintenance. To maintain the community, we will need to engage with developers to evaluate the sustainability of development efforts, and engage with users to evaluate the effectiveness of the platform in improving access to bioinformatics and genomic analysis.

and the intended use and long-term value of the data and appropriate metadata standards. Archiving practices must be responsive to the requirements and priorities of individual health jurisdictions, but we recommend that wherever possible, agencies prioritize keeping data easily searchable and shareable.

Make bioinformatics pipelines fully open-source and broadly accessible. Currently, commercial software can provide off-the-shelf bioinformatics capabilities to laboratories with limited in-house capacity. However, licensing proprietary software can be prohibitively expensive, especially in low- and middle-income settings. Though they are perhaps less obvious, proprietary software also has limitations in high-income settings; although it may be economically accessible, using proprietary software may reduce transparency about how data are processed, and it limits customization of bioinformatics pipelines.

To facilitate broad access to standardized bioinformatics across income settings, we recommend developing and maintaining a deployment platform of open-source pipelines for bioinformatics assembly and genomic analysis. We describe a model of this deployment platform in greater technical detail in Box 3. Within this ecosystem, we recommend that bioinformaticians use open-source software packages within pipelines, and that they deploy full pipelines openly. Pipelines should output to common, non-proprietary file formats, and bioinformaticians should build them transparently in an environment that supports user feedback and issue tracking. To ensure that limited informatic training is not a barrier to use, frequently used reference pipelines, such as those used for molecular surveillance of foodborne pathogens, should be accessible via web-based entry portals with graphical user interfaces.

We note that access to standardized bioinformatics does not mean limiting the number of workflows available. Rather, it means ensuring that we build software and workflows upon widely accepted standards in a way that is transparent and auditable by the community. If interoperability between tools and openness of the entire system to sharing and review are prioritized, we believe that a balance point will be reached at which there are sufficient tools and workflows to support the analyses public-health agencies want to perform, without leading to a proliferation of redundant tools and workflows (Box 3).

Develop modular pipelines for data visualization and exploration. A large proportion of genomic data interpretation relies on data visualization, such as the creation of phylogenetic trees. However, the current process for making and refining these visualizations is inefficient. Genomic data are frequently separated from epidemiologic data, and most public-health bioinformaticians will not have access to demographic and exposure information for the individual who is the source of the data. This means that bioinformaticians cannot easily analyze epidemiologic and genomic data jointly to create integrated visualizations.

Additionally, visualization pipelines typically run as a monolithic series of computations that start from raw sequencing reads and end with a single image, not a genomic data object that can be visualized

Box 4 | The reproducible bioinformatics environment

Here we outline how versioning, containerization, auditability, validation and workflow management software contribute to reproducibility.

Versioning. At its heart, versioning allows anyone editing code or a dataset to document and track the changes they make. Documentation allows other developers or data users to understand what has changed and why, and tracking enables individual changes to be rolled back if necessary. Versioning also enables bioinformaticians to develop newer generations of pipelines, or customized pipelines, and test them, without inhibiting access to stable reference pipelines. Taken together, versioning creates a transparent environment where developers can make, find and fix mistakes; where users can access validated pipelines without preventing developers from updating them; and where customized pipelines and reference pipelines can coexist. We recommend that data curators use version control to track and document changes to reference datasets, and that bioinformaticians version both component software programs and whole pipelines.

Containerization. We recommend that bioinformaticians and software developers containerize full pipelines as well as software packages that are used outside of pipelines. Containerization increases the reproducibility of analyses by making it possible for one user to run the same pipeline in the exact same computing environment as someone else. This consistency in the computing environment limits problems in which missing dependencies (additional pieces of software that must be installed so that the desired software package can run), or differences in versioning of dependencies, change the way a pipeline runs. We recommend releasing containerized reference pipelines as versioned generations, following a stable release cycle. All generations of a pipeline should be concurrently hosted on the platform to ensure historical compatibility of bioinformatics analyses. Having these different generations of pipelines hosted together also allows developers to benchmark pipelines side by side. This ability to compare workflows systematically within the same environment is critical to ensuring that bioinformatics assays remain valid even as bioinformaticians update them. Bioinformaticians working in public health have already begun to containerize useful software, such as the library of Docker builds maintained by the State Public Health Bioinformatics group (https://github.com/StaPH-B/docker-builds). This effort could be developed further, or bioinformaticians could also use other containerization projects, such as BioContainers (https://github.com/BioContainers/specs), Bioboxes (http://bioboxes.org/) or FlowCraft (https://github.com/assemblerflow/flowcraft).

Auditability. We recommend recording the processes that a pipeline has performed. Pipelines and workflows could automatically generate reports describing the name and version of each software component, as well as the data inputs and settings. Additionally, when a user runs a pipeline, the pipeline should automatically store intermediate files in standardized formats, such as FASTA, CSV or JSON. Having access to these intermediate files supports troubleshooting, as they can reveal the presence of discrepancies and where those were introduced.

Validation. Although validation datasets exist (see ref. [16]), we recommend developing additional structured validation criteria for bioinformatics assembly pipelines. We imagine that agencies at higher levels of jurisdictional authority would be responsible for developing validation metrics, because these standards would apply to a broad range of agencies. Additionally, we suggest that agencies perform end-to-end proficiency testing of whole-genome sequencing protocols, including both the laboratory and bioinformatics portions of the assay. Finally, bioinformaticians, or anyone releasing pipelines to the deployment platform, should clearly communicate which pipelines have been formally validated.

Workflow management. One of the best strategies for writing reproducible and auditable pipelines is to design them as automated workflows. Although a pipeline can be written as a single script, specifying pipelines in workflow languages inherently documents the steps that the pipeline will follow, as well as expected data inputs and outputs. To maximize the portability of pipelines across platforms, workflows could be written in Common Workflow Language (https://www.commonwl.org), which would allow them to run on various deployment platforms such as Arvados (https://doc.arvados.org/) and Terra/FireCloud (https://support.terra.bio/hc/en-us), and eventually also on Galaxy (https://usegalaxy.org/). Alternatively, pipelines could be written with other workflow systems, such as Snakemake (https://snakemake.readthedocs.io/en/stable/) or Nextflow (https://www.nextflow.io/). Although they are potentially less portable, these workflow systems have high uptake in biology and may be more familiar to developers in public health.

and explored in multiple ways. This image is generally shared over email, and manually annotated with epidemiologic data. Highly collaborative teams seeking to integrate their genomic and epidemiologic interpretations may repeat this cycle of generating images and then annotating images many times over; this is time consuming and potentially error prone, and it may limit which analyses can be performed, reducing public-health utility.

We recommend taking a more functional, modular approach. Firstly, analytic and visualization pipelines should be separated from assembly pipelines. This separation allows genome assembly to occur on lower-security scientific computing servers in the absence of epidemiologic data. The separation also provides an added benefit that if a bioinformatician wishes to rerun an analysis, they do not need to redo the genome assembly. After assembly, bioinformaticians could join epidemiologic data housed on secure servers with the assembled genomes. If APIs are used to source data, different levels of security authorization could be required to access different components of epidemiologic data. Notably, for data joining to work, structure and consistency provided by data models and ontologies will be needed. Finally, rather than exporting a single image, analytic pipelines could export data objects for interactive visualization in browser-based portals, increasing epidemiologists' and bioinformaticians' capacity to explore the data together.

Interpreting genomic data is not always intuitive, which can make communicating findings to multidisciplinary public-health teams challenging. To improve data interpretation, we recommend developing analytic tools further, so that they properly account for uncertainty in the sampling process, and developing new ways to convey uncertainty within genomic data visualizations. The widespread use of genomic data in public health is relatively new, and many public-health practitioners do not have a background in

Box 5 | Transitioning to the cloud

General concerns. To date, issues with process, compliance and acquisition of cloud services by governmental agencies at all levels of jurisdictional authority have hindered the adoption of cloud computing in public health. However, as cloud services become increasingly feasible for government agencies to access, we expect their utility to increase.

Shifting to cloud computing converts capital expenses to operational expenses, which hopefully will make it easier to spend money on computing resources. Although many agencies likely understand the traditional capital and operational expenditures associated with purchasing and maintaining servers, probably fewer have a good understanding of how cloud computing operational expenditures compound, and how to install necessary controls on them. To ease expenditure concerns and smooth adoption, we recommend that public-health programs receive training on how cloud operational expenditures work, how to install controls and how to train users who are purchasing resources.

Finally, as research moves to the cloud, agency-specific data and patient privacy policies will need to be updated to permit cloud-based computing. We recommend that public-health agencies communicate openly with cloud computing providers, such that cloud services can be tailored to the needs and standards of public-health agencies.

Cloud computing in low- and middle-income countries. Erratic internet connectivity and limited bandwidth could hinder the development of cloud-based bioinformatics in low- and middle-income countries. Intermittent connectivity is less of an issue for running bioinformatics analyses, as once the data are in the cloud, genome assembly can proceed easily. Rather, internet interruptions impact the movement of data between local machines and the cloud.

For example, during most of the 2014–2016 outbreak of Ebola virus in West Africa, there was no method for performing offline bioinformatics analysis of sequencing reads from the portable MinION sequencing devices used to collect data in the field. This meant that scientists needed to upload reads to the cloud for analysis, using a system that engineers designed for stable internet connections. In West Africa, the internet connectivity was generally insufficient to easily complete the data upload process. This meant that raw sequencing reads were frequently uploaded over internet hotspots using the 3G mobile network[17]. Though tenable in an emergency situation, this solution is likely unsustainable for large genomic surveillance programs.

Although interruptions to internet connectivity and changes in bandwidth may occur unpredictably, connectivity and bandwidth are usually sufficient for data uploading when averaged over longer time periods (D. Park, personal communication, 2 May 2020). Thus, to support cloud-based bioinformatics in Africa, we likely do not need completely different pipelines. Rather, we should develop fault-tolerant mechanisms for uploading raw genomic data to the cloud. Currently, the African Centre of Excellence for Infectious Diseases at Redeemer's University, Nigeria, uses a cloud-based version of the Broad Institute's viral-ngs pipeline[18]. Although the assembly pipeline itself is the same as that used in the USA, bioinformaticians have made the data upload process for Africa more resilient and persistent. During upload, the data are divided into smaller chunks, and only portions are uploaded at a time. This means that, when connectivity interruptions occur, only a small part of the whole task must be restarted. The workflow for data upload is also persistent, automatically restarting if the internet connection is interrupted, and attempting to upload data over an entire week. This system works, and has supported regular assembly of Lassa virus genomes[19] and those of other pathogens over the last few years.

genomic epidemiology. Thus, researchers must ensure that data exploration and visualization tools effectively capture and convey uncertainty to experts and non-experts alike.

Improve the reproducibility of bioinformatics analyses. As often occurs in academic settings, public-health programs routinely use similar, but distinct, pipelines for bioinformatics analysis. Although most pipelines use a relatively narrow suite of open-source software programs, the lack of standardization across bioinformatics pipelines affects the comparability of data and results across agencies.

Sequencing assays in public health must be sufficiently robust and reproducible to meet government-regulated standards. This need for stable software and reproducible analyses should drive how bioinformatics pipelines are developed, maintained, hosted and tested. To meet this need, we recommend using version control to manage datasets and pipelines, containerizing code and requirements, auditing pipelines, using workflow management software and developing validation criteria for assessing bioinformatics assembly against known standard datasets (Box 4).

Utilize cloud computing to improve the scalability and accessibility of bioinformatics analyses. As the scope of genomic surveillance grows, so too will the volume and complexity of data generated during routine public-health laboratory operations. For many public-health institutions, the assembly and analysis of next-generation sequence data already depends on advanced computing infrastructure for data capture, analysis and storage. To better support the current needs, as well as to plan for the future, we recommend developing the public-health bioinformatics ecosystem as a cloud-based system. We imagine that the cloud-based system would be hosted centrally, probably by a federal public-health agency, which would reduce the number of high-performance computing environments needed to support broad access to bioinformatics. That way, not every institution would have to purchase server hardware nor pay the highly remunerated workforce necessary to maintain a high-performance computing cluster. Instead, smaller agencies could pay only for their usage of the cloud-based ecosystem, and even this could be reduced if computing were entirely centrally funded. Agencies could manage costs more efficiently due to the inherent elasticity of cloud computing. Computing power could be scaled up in times of high demand, such as outbreaks, and scaled down when demand is low to reduce costs.

Centralized management of a broadly accessible resource would also allow agencies in smaller jurisdictions, or in low- and middle-income countries, to support sophisticated bioinformatics capabilities without incurring substantial capital or operational expenditures. Broadening the access to bioinformatics could help build capacity within small frontline public-health agencies, thereby reducing lag times during outbreak response. Broader access would also enable smaller agencies to investigate priority diseases at the local level.

In addition to scalability, accessibility and potential economic benefits, a cloud-based bioinformatics ecosystem could also improve the reproducibility of bioinformatics analysis. To run on the cloud, code should be containerized (Box 4). If most agencies and programs use the same pipelines, their results will be more comparable.

Although cloud-based computing holds great potential for public health, we would be remiss if we did not mention one formidable obstacle: connectivity issues in low- and middle-income countries (Box 5).

Support new infrastructure and software development demands with an expanded technical workforce. Tomorrow's public-health workforce should include new technical specialties. To support the computational infrastructure necessary for broad-scale public-health genomics, programs could benefit from personnel with expertise in managing high-performance computing infrastructures, storage engineers who manage databases and networks, and software developers. To support the analysis of growing amounts of complex data, public-health agencies could benefit from additional bioinformaticians, genomic epidemiologists and data scientists.

Attracting this workforce may be challenging. Lower compensation than in the private sector, lack of access to newer technologies and the different culture of working within a government agency could prevent computationally oriented personnel from pursuing careers in public health. Emphasizing the ability to improve lives may increase recruitment, however.

Beyond recruitment, public-health programs should consider retraining as a way to build this workforce. Increasingly, laboratory microbiologists are pivoting towards more bioinformatics-heavy roles, often by learning these new skills on their own. With their incredible wealth of knowledge about the upstream sequencing process, former microbiologists have a unique perspective that could improve troubleshooting and evaluation of bioinformatics analyses. Bioinformatics training programs now exist across resource settings—for example, H3ABioNet (https://www.h3abionet.org/), ELIXIR-Tess (https://tess.elixir-europe.org/), GOBLET (https://www.mygoblet.org/) and Australian BioCommons (https://www.biocommons.org.au/training)—and these could serve as models for public-health bioinformatics training. Although public-health programs should design multiple courses tailored to different skill levels, possible topics could include command-line interfaces and common platforms for bioinformatics analysis, interpreting quality control metrics for whole-genome sequence data, bioinformatics methods for genome assembly, and the theory and practice of comparative genomic analysis.

Once such a workforce is developed, public-health agencies will also need to retain its members. Currently, many agencies lack formal job descriptions specific to computational disciplines, competency and assessment criteria, and mechanisms for computational personnel to advance into leadership roles. To sustain a computational workforce, public-health agencies should create clear descriptions of the disciplines and job series for bioinformaticians, data scientists and software engineers within public health.

Retaining computational personnel will probably be more challenging in low- and middle-income countries, and we expect that the recommendations above will be insufficient. As discussed by Folarin and colleagues[14] in relation to retaining African scientists in genomic research, retention will likely require coordinated governmental support, sustained funding and infrastructure development. Public-health practitioners in low- and middle-income countries will know best how to approach building and retaining capacity, and we defer to their knowledge and experience. We simply note that understanding how to retain computational personnel in low- and middle-income settings is critically important to developing an informatic infrastructure that can work across all income settings.

Improve the integration of genomic epidemiology with traditional epidemiology. Neither epidemiologic case data nor pathogen genomic data are as powerful on their own as they are when integrated and analyzed together in a timely and actionable manner. From a technical perspective, this integration will require more sophisticated databasing approaches, including programmatic data sourcing and merging that respect security levels, use of ontologies to standardize data reporting formats for both surveillance data and genomic data, and machine-learning methods for data classification, tagging and cleaning.

Even with the necessary technical requirements in place, however, effective integration of epidemiologic and laboratory data will also require frequent and open communication between surveillance epidemiologists and bioinformaticians. We believe that this communication could be improved if bioinformaticians and epidemiologists could, to a certain degree, speak each others' languages. We recommend that public-health agencies train surveillance epidemiologists in the basics of interpreting genomic data and, likewise, teach bioinformaticians the basic concepts of epidemiology. For example, a course for surveillance epidemiologists could clarify the applicability of genomics to their work, describe how genomic data are generated and discuss possible epidemiologic interpretations from comparative genomic analyses. Similarly, bioinformaticians may not always understand how epidemiologic questions and study design shape sample selection for sequencing. Thus bioinformaticians may benefit from courses that describe epidemiologic study designs, common analytic techniques in epidemiology and principles of public-health surveillance.

We also recommend creating integrated teams that include genomic epidemiologists, bioinformaticians and surveillance epidemiologists. We imagine that harnessing this expertise across disciplines will strengthen epidemiologic interpretations, as

Box 6 | Learning from the success of BioNumerics and PulseNet

Large, diverse genomic datasets from many groups are greater than the sum of their parts, and their utility has built momentum for greater data sharing within some sectors of public health. One of the best examples is PulseNet, a laboratory-based surveillance network for foodborne bacterial disease. Data sharing within PulseNet occurs along trusted channels, built on memorandums of understanding with each of the collaborating partners. These memorandums describe how data will be shared, with whom and at what granularity, ensuring compliance with state and federal law.

Software also plays a major role in supporting PulseNet. BioNumerics (Applied Maths/bioMérieux) provides an intuitive bioinformatics analysis toolkit with a graphical user interface, which has enabled frontline public-health agencies to analyze molecular data from multiple pathogens. This ability gives agencies greater autonomy to investigate diseases that are priorities at the local level. BioNumerics also makes data sharing easy by integrating sharing mechanisms into the software. A user can add detailed and complete data about a sample to a local BioNumerics database, and then share subsets of that data with PulseNet via easy interaction within the BioNumerics. In this way, BioNumerics acts as a national database with specimen and process tracking and within-network data sharing.

BioNumerics has yielded many benefits that researchers should strive to maintain. But we must also recognized its limitations, and work to improve upon them. Most critically, BioNumerics is not open source, and licensure costs can be prohibitively expensive for small institutions and institutions in low- and middle-income countries. Additionally, its lack of modularity limits options for custom development and expansion of pipelines. Thus, we still recommend moving toward a fully open ecosystem. However, much has worked within the BioNumerics environment, and the community must strive to maintain those qualities during the transition.

Box 7 | The COVID-19 pandemic as a case study

At the time of writing, five months had passed since the first SARS-CoV-2 genome sequence was released. During that time, researchers and public-health agencies from 80 countries submitted thousands of SARS-CoV-2 genomes into genomic databases such as GISAID, ENA and NCBI GenBank. This is an incredible amount of data generated by a large number of organizations. Given the large number of sources of data, systematic ways are needed to ensure that data are consistent across submitting agencies—the field needs a widely adopted data model. In the absence of a data model, sequence data and relevant metadata will be variably present and formatted. For example, we have found that some SARS-CoV-2 sequences give only the year of collection, other sequences have year and month information, and others have the year, month and day of sample collection. In other instances, we have found sequences for which the submitting author entered the date of sequencing as the sample collection date, which is likely to prevent case identification. Although the reasons for this heterogeneity are understandable, the variability impacts analyses of the sequence data. For SARS-CoV-2, which has circulated for only a few months and therefore lacks a strong evolutionary rate signal, knowing the full date of collection, including year, month and day, is critically important for inferring temporally resolved phylogenies. Currently, standardizing these data can require direct communication between data submitters, which is time consuming. Although some degree of manual curation will always be necessary, unified data models could make curation easier.

When it comes to the sequences themselves, bioinformaticians will likely have assembled the genomes using different bioinformatics pipelines. Although the impact of this variation is hard to quantify precisely, we are relatively certain that variations in pipeline parameters, such as the minimum read depth necessary to make a base call, mean that different pipelines could produce slightly different sequences given the same raw reads. This underscores the need for an open and accessible bioinformatics ecosystem, which would allow bioinformaticians to share pipelines and evaluate different pipelines to understand why assembled genomes may differ slightly.

Finally, phylogenetic analysis of SARS-CoV-2 genomes has helped to elucidate the epidemiology of the pandemic, from initially providing evidence of human-to-human transmission to exploring geographic patterns of transmission. Although these inferences are valuable, there is also a considerable risk of misinterpretation. Some large outbreaks are not well represented in the sequence data, whereas others are incredibly over-represented. These differences in sampling intensity can affect phylogeographic inference and our understanding of transmission histories, but non-experts might not know how. The value of the knowledge can be gained from these data, as well as the risk of misinterpretation, highlight the importance of integrating genomic and traditional epidemiology within public health, as well as improving the ways in which we communicate uncertainty.

public-health officials will make inferences from multiple data sources, and the strengths and weaknesses of each data source will become clearer.

Develop best practices to support open data sharing. In an interconnected world where disease transmission occurs across borders, environments and species, the best surveillance system would support data sharing across institutions and agencies, both within and between countries. Ideally, all genomic data and non-identifiable metadata would be shared openly between all agencies, with data release occurring rapidly after data generation, once data are in a reasonably reliable draft form. Although harder to share, personally identifiable information (PII) can be critically important to understanding an outbreak. We recommend sharing PII along secure and trusted channels to the extent that it is important for understanding disease dynamics and guiding public-health responses.

Although we advocate for the described degree of openness as a best practice, we recognize that these standards would be nearly impossible to implement within our current public-health system. Some diseases, such as HIV, will require stricter constraints on data sharing for as long as they remain stigmatizing. Other infections are rare, allowing one to rapidly triangulate from non-identifiable data to PII. Finally, although public-health programs rightfully must follow rules that govern how PII are shared, these regulations often make data sharing convoluted, because definitions of PII vary by disease incidence and geography, and laws governing the use, storage and transmission of PII vary by jurisdiction.

In order to develop a data sharing system that functions well for public health, we think that data sharing needs to be easy to do, so that it is not a burden; occur along trusted channels; and be granular, so that access to different levels of data can be filtered based on security and legal constraints. We emphasize that the development of increased data openness in public health cannot be all or nothing; if it is, we will simply end up with a system in which sharing is limited. Instead, we should identify consistent small steps that programs can take to improve the openness of data, with the hope that open data and integrated databases improve surveillance and outbreak response sufficiently to warrant their continued development and maintenance (Boxes 6 and 7).

Our vision of a potential software ecosystem

Given our proposals, and the software tools that currently exist, we imagine that our proposed system would be highly modular, with genomic assembly and data processing separated from the genomic analysis and visualization processes. Splitting these processes will maintain efficiency while allowing flexibility, enabling many different analyses to be performed without the need to rerun assembly pipelines. Importantly, separating the assembly and the analytic processes will also ensure that output from the assembly pipelines is archived, an important extension to current archival practices, which focus primarily on storing raw sequencing reads. The primary pieces of this ecosystem would be databases, APIs, pipelines and scripts that move data around (Fig. 1).

Conclusion

The shift toward extensive use of pathogen whole-genome sequencing represents a turning point for public-health agencies; programs must pivot to accommodate a new data source that provides increased resolution for understanding disease dynamics, but requires different tools and a changing workforce to support. Now is the time to build community and consensus, to invest in developing a system that is broadly accessible and that will work for years to come.

Our recommendations provide a starting point for these discussions. The efforts to realize an open ecosystem for public-health bioinformatics will be guided and supported by the Public Health Alliance for Genomic Epidemiology (PHA4GE), an organization that we, along with many others from the public-health bioinformatics community, launched in 2019. PHA4GE is a global coalition that is actively working to establish consensus standards, document and share best practices, and improve the availability of critical bioinformatics tools and resources. Through its work, we hope to see greater openness, interoperability, accessibility and reproducibility in public-health bioinformatics.

Received: 26 March 2020; Accepted: 8 May 2020;
Published online: 11 June 2020

References

1. Armstrong, G. L. et al. Pathogen genomics in public health. *N. Engl. J. Med.* **381**, 2569–2580 (2019).
2. Centers for Disease Control and Prevention. Laboratory surveillance for wild and vaccine-derived polioviruses, January 2002–June 2003. *Morbid. Mortal. Wkly Rept.* **52**, 913–916 (2003).
3. Doyle, R. M. et al. Direct whole-genome sequencing of sputum accurately identifies drug-resistant *Mycobacterium tuberculosis* faster than MGIT culture sequencing. *J. Clin. Microbiol.* **56**, e00666–18 (2018).
4. Lu, R. et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565–574 (2020).
5. Andersen, K. G., Rambaut, A., Lipkin, W. A., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nat. Med* **26**, 450–452 (2020).
6. Eden, J.-S. et al. An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. *Virus Evol.* **6**, veaa027 (2020).
7. Lu, J. et al. Genomic epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell* **181**, 997–1003.e9 (2020).
8. Zehender, G. et al. Genomic characterisation and phylogenetic analysis of SARS-COV-2 in Italy. *J. Med. Virol*. https://doi.org/10.1002/jmv.25794 (2020).
9. Bedford, T. et al. Cryptic transmission of SARS-CoV-2 in Washington State. Preprint at https://doi.org/10.1101/2020.04.02.20051417 (2020).
10. Fauver, J. R. et al. Coast-to-coast spread of SARS-CoV-2 during the early epidemic in the United States. *Cell* **181**, 990–996.e5 (2020).
11. Gwinn, M., MacCannell, D. & Armstrong, G. L. Next-generation sequencing of infectious pathogens. *JAMA* **321**, 893–894 (2019).
12. Dooley, D. M. et al. FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *NPJ Sci. Food* **2**, 23 (2018).
13. Wilkinson, M. D. et al. The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
14. Folarin, O. N., Happi, A. N. & Happi, C. T. Empowering African genomics for infectious disease control. *Genome Biol.* **15**, 515 (2014).
15. Fielding, R. T. & Taylor, R. N. *Architectural styles and the design of network-based software architectures* Doctoral dissertation, Univ. California, Irvine (2000); https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm
16. Timme, R. E. et al. Benchmark datasets for phylogenomic pipeline validation, applications for foodborne pathogen surveillance. *PeerJ* **5**, e3893 (2017).
17. Quick, J. et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232 (2016).
18. Park, D. et al. https://zenodo.org/record/3509008 (18 October 2019).
19. Siddle, K. J. et al. Genomic analysis of Lassa virus during an increase in cases in Nigeria in 2018. *N. Engl. J. Med.* **379**, 1745–1753 (2018).

Acknowledgements

We are grateful to all of the public health professionals who took the time to participate in these interviews and discuss their experiences with us in depth; without them, this work would not have been possible. We thank all the participants in the Open Public Health Bioinformatics Meeting for contributing to this discussion, and in particular D. Aanensen, J. Carriço, R. Neher and D. Park for their edits and suggestions to this paper. This work was supported by a grant from the Bill and Melinda Gates Foundation to the CDC Foundation. A.B. is supported by the National Science Foundation Graduate Research Fellowship Program under grant no. DGE-1256082. T.B. is a Pew Biomedical Scholar and is supported by NIH R35 GM119774-01. The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Diseases Control and Prevention.

Competing interests

The authors declare no competing financial interests.

Additional information

Correspondence should be addressed to D.R.M.

Peer review information Hannah Stower was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.