nature microbiology

Perspective

https://doi.org/10.1038/s41564-022-01233-6

The potential of genomics for infectious disease forecasting

Received: 1 March 2022

Accepted: 18 August 2022

Published online: 20 October 2022

Check for updates

Jessica E. Stockdale, Pengyu Liu and Caroline Colijn ✉

Genomic technologies have led to tremendous gains in understanding how pathogens function, evolve and interact. Pathogen diversity is now measurable at high precision and resolution, in part because over the past decade, sequencing technologies have increased in speed and capacity, at decreased cost. Alongside this, the use of models that can forecast emergence and size of infectious disease outbreaks has risen, highlighted by the coronavirus disease 2019 pandemic but also due to modelling advances that allow for rapid estimates in emerging outbreaks to inform monitoring, coordination and resource deployment. However, genomics studies have remained largely retrospective. While they contain high-resolution views of pathogen diversification and evolution in the context of selection, they are often not aligned with designing interventions. This is a missed opportunity because pathogen diversification is at the core of the most pressing infectious public health challenges, and interventions need to take the mechanisms of virulence and understanding of pathogen diversification into account. In this Perspective, we assess these converging fields, discuss current challenges facing both surveillance specialists and modellers who want to harness genomic data, and propose next steps for integrating longitudinally sampled genomic data with statistical learning and interpretable modelling to make reliable predictions into the future.

The development of high-throughput sequencing has transformed biology and medicine. It is now possible to analyse thousands of genomes in a single study, and sequencing-derived technologies have had tremendous impact: detecting alleles associated with cancer or genetic disorders, characterizing and detecting antimicrobial resistance (AMR), studying microbial diversity and more[1]. Sequence data present a high-resolution view of the processes of diversification and adaptation, the origins of phenotypes of interest and the myriad ways that diversity may be acquired, lost or maintained. Phylogenetic tools allow inference of patterns of ancestry from observed diversity, and sampling and sequencing through time reveal how measurably evolving organisms have changed and adapted on observable timescales. When this change has happened in the presence of selection, environmental variation, genetic drift and population bottlenecks, sequencing technology and temporal sampling provide the opportunity to learn about evolution.

Pathogen diversification presents health challenges, with the rising burdens of AMR being a clear example. Since antibiotics were first introduced, clinical resistance has consistently followed the introduction of new antimicrobials within one to two decades[2]. Viral evolution is rapid, and treatment of fast-evolving infections such as human immunodeficiency virus (HIV) is challenging due to the speed at which some viruses can acquire resistance[3]. Influenza viruses evolve through patterns of antigenic drift and periodic antigenic shift; seasonal vaccines need to be updated regularly, and pandemic strains can emerge repeatedly[4]. The issue of pathogen diversification has come to the forefront during the coronavirus disease 2019 (COVID-19) pandemic, with the continued emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variants of concern driving epidemic waves[5]. Even in slowly evolving pathogens such as *Mycobacterium tuberculosis*, extensively resistant variants have been reported worldwide[6]. In *Plasmodium falciparum*, a eukaryotic organism that causes

Department of Mathematics, Simon Fraser University, Burnaby, British Columbia, Canada. ✉e-mail: ccolijn@sfu.ca

human malaria, resistance to antimalarial agents is a key factor driving global malaria increases[7].

There are increasing efforts to compile genomic data in publicly accessible databases, with a focus on resistance. With large bacterial sequence datasets, researchers have characterized recombination pathways[8], capsule switching and resistance acquisition following human intervention in *Streptococcus pneumoniae*[9], identified resistance determinants in *M. tuberculosis* and *Escherichia coli*[10,11] and characterized global patterns of cholera dissemination[12], to name a few. Recently, over 2,000 whole-genome sequences of *Neisseria gonorrhoeae* were analysed alongside epidemiological data, revealing a novel resistant clone and transmission among distinct contact networks[13]. Researchers have identified mutations that confer drug resistance in HIV[14] and hepatitis C virus[15] and key differences in within- and between-host evolution that affect the development of resistance[16]. Many COVID-19 studies leveraged large volumes of sequence data: more than six million SARS-CoV-2 genomes were analysed to identify mutations associated with transmissibility of the virus[17]. Since the conception of 'viral phylodynamics' in 2004[18], models can estimate underlying parameters using likelihoods for phylogenetic trees, linking mechanistic models of diversity with genomic data. Estimated parameters can be used in forward-time models to make predictions of the relevant population dynamics. This approach effectively summarizes the information in a set of pathogen sequence data as one or several real parameters.

In contrast, models used for infectious disease forecasting often cannot incorporate pathogen diversity and cannot typically be compared with genomic data. These models include the susceptible–infectious–recovered compartmental model and its myriad extensions (for example, latent periods, age structure and vaccination status). Although recent work links birth–death and coalescent phylogenetic models to compartmental models and projections[19–22], many modelling frameworks used for forecasting disease do not yet lend themselves to modelling diverse pathogens. For example, a report[23] estimates that by 2050, AMR will cost up to US$100 trillion and cause 10 million deaths per year, based on an assumption that all bacterial infections will be resistant. But in many of the most prevalent bacteria causing human disease, resistance has remained at stable intermediate frequencies for many years[24,25]. Modelling even this single fact about bacterial diversity (resistant and sensitive types can coexist for long periods) has proved challenging[25], but if models do not correctly describe standing, stable diversity, they have poor prospects for making good forecasts into the future.

In this Perspective, we outline the need and opportunity for stronger links between forecasting and genomic data. Sequencing technologies have matured to the point where sampling in a consistent manner over time is feasible, and this gives us the opportunity to observe the evolution of our most important pathogens in response to our interventions. If we could incorporate these data into predictive models—building, testing and refining them against high-resolution data on evolution through time—we would stand a much better chance of assessing risks of immune escape, AMR and other evolutionary changes, and mitigating these risks. We describe several recent efforts in this direction, the availability of relevant data and the remaining challenges. We call on modellers and genomics experts to create the data, models, benchmarking and refinements that will be required to bring genomic data together with forecasting efforts.

Rapid increase in genomic data

Large volumes of pathogen sequence data have been collected and made available online (Table 1). The Genomes Online Database (GOLD)[26] links studies and metadata, sourced from the National Center for Biotechnology Information (NCBI), the Department of Energy Joint Genome Institute and others. Other databases host large amounts of viral sequence data (for example, NCBI's Virus portal[27] and the Global Initiative on Sharing All Influenza Data [GISAID][28]). The Pathosystems Resource Integration Center (PATRIC) database[29] collates genomes of pathogenic bacteria, with available year and country data and some antibiotic resistance information. The quantity of documented sequences has been exponentially increasing over time[30].

Table 1 | A selection of sequence databases containing pathogen genomes

Database/ project	Collection or aim	Reference
GOLD	>460,000 organisms (primarily pathogenic), collated from multiple sources	26
NCBI Virus	>1.3 million viral sequences, spanning RNA, DNA and unclassified viruses	27
GISAID	>365,000 influenza and >10.8 million SARS-CoV-2 sequences	28
PATRIC	>570,000 genomes of pathogenic bacteria	29
Global Pneumococcal Sequencing Project	>26,000 pneumococcal genomes	https://www.pneumogen.net/gps/
Los Alamos HIV sequence database	>980,000 records and >16,000 complete HIV genomes	95

Quantities are correct as of 17 May 2022.

There are a number of projects providing genomic data for the detection, comparison and study of AMR genes and isolates. Major databases include the Comprehensive Antibiotic Resistance Database[31], MEGARes[32], DeepARG[33], and the broader-purposed Uniprot[34]. Knowledge of the emergence and origins of AMR is essential in preventing and mitigating its damages. However, large collections of AMR genes, with differing sampling strategies, without the organisms' context and without information about antibiotic-sensitive counterparts, are not directly amenable to forecasting and modelling pathogen ecology and evolution. Continual sampling as part of surveillance programmes may close this gap to some extent, particularly if sampling also includes non-resistant and 'background' isolates and organisms. In this vein, the European Antimicrobial Resistance Surveillance Network project (https://www.ecdc.europa.eu/en/about-us/partnerships-and-networks/disease-and-laboratory-networks/ears-net) is focused on AMR surveillance, collecting data from invasive isolates originating from national surveillance programmes and laboratory networks. The US Antibiotic Resistance Laboratory Network (https://www.cdc.gov/drugresistance/ar-lab-networks/domestic.html) spans 50 states and Puerto Rico, and reports AMR to the US Centers for Disease Control and Prevention, which runs the Antibiotic Resistance Solutions Initiative. The World Health Organization's (WHO's) Global Antimicrobial Resistance Surveillance System (https://www.who.int/initiatives/glass) is promoting the development of additional national surveillance systems to collect, analyse and share data.

There are ambitious long-term sequencing projects underway, some linked to surveillance programmes. The Comprehensive Resistance Prediction for Tuberculosis: An International Consortium project (http://www.crypticproject.org/) is sequencing 100,000 genomes for tuberculosis from five continents, and both England and the United States use routine whole-genome sequencing for tuberculosis. The Wellcome Trust Sanger Institute's Parasites and Microbes programme has an ongoing commitment to sequencing a range of organisms and making data available (Table 1). There are many clinical, reference and public health laboratories around the world that have stored isolates over many years; these isolate collections could be sequenced. In all, including existing datasets, upcoming projects and the decreasing cost

https://doi.org/10.1038/s41564-022-01233-6

of sequencing existing collections of isolates, there are rich opportunities to build the data to capture, at a high level of resolution, ongoing pathogen evolution.

Epidemiology and genomic data

The rapid accumulation of genomic data has provided insight into epidemiological and evolutionary processes and stimulated the development of a number of methods[18]. These have been applied to inferring epidemiological parameters, investigating transmission patterns[35], determining the spatial, temporal and zoonotic origin of pathogens[36], understanding acquisition and transmission of AMR[37], and modelling fitness[38].

Pathogen genomic data encode information for inferring epidemiological parameters including the basic and effective reproduction numbers and the effective population size through time. Several inference methods have been developed for large-scale models with relatively sparse sampling, for example, to estimate the basic reproduction number using Bayesian inference with a birth–death model for HIV-1 virus in Switzerland[39], and with a structured coalescent model for SARS-CoV-2[40]. Recent extensions to these approaches have allowed for differences between lineages, inter-strain interactions and geographic movements[41–43]. Multi-type branching process models allow for rapidly evolving or co-circulating pathogens and host populations with heterogeneous contact structures. Tree comparison approaches estimate parameters by comparing phylogenetic trees from simulations with those from data, via approximate Bayesian computation (ABC)[44] or via mathematical representations of phylogenetic trees[45–47].

Efforts have also been made to reconstruct outbreak transmission trees from genomic data in outbreak settings with dense sampling[48,49]. Genomic data are also used to understand contact networks, for example, using ABC for HIV genomic data to estimate structural parameters of contact networks[50], and to identify transmission risk factors, for example, through clustering or viral diversification rates[51]. At much larger scales, genomic data have been used in influenza virus research to predict evolutionary change[38], with the potential to inform vaccine design. This is enabled in part by routine collection of influenza sequences, linked to geographic and epidemiological information.

Pathogen genomic data can be more informative for predictive models when linked to metadata. Epidemiological models exploiting linked genomic data have been developed for a limited number of pathogens due to the availability of metadata. Methods that unify classic infectious disease compartmental models and population dynamics from genomic sequences[19,52] have gained popularity, as they allow description of phylogenetic clustering patterns in addition to epidemiological parameter estimation. Methods combining epidemiological and genomic data have been used to reconstruct early transmission trees of foot-and-mouth disease outbreaks[53] and to infer likely infection times and heterogeneity in infection[54]. With locations of sampled genomic sequences, phylogeographic methods help characterize the emergence of a pathogen, identify importation and local circulation, and evaluate factors driving transmission. With a structured coalescent model and Bayesian inference, genomic data and their sampling locations have been used to reconstruct transmission histories, migration patterns and outbreak origins[43,55]. Although not directly predictive, this has policy applications: it has been shown that many regional outbreaks of SARS-CoV-2 virus (in New York City, Israel and others) were initiated by multiple introductions, highlighting the highly porous nature of borders[56,57].

Currently, these methods have been mainly retrospective or descriptive. We propose that it is possible, and it is time, to develop models with genomic data that produce results relevant to prediction. Data availability to support this effort is improving. Linkage of genomic data to metadata and other epidemiological information remains limited, although it would render genomic data far more informative and improve forecasting efforts[58]. Estimation of model parameters and incorporation of some epidemiological structure[19,22] has been a clear step towards prediction, and strengthening models with additional epidemiological data would allow genomic methods to be more effectively linked with current forecasting efforts.

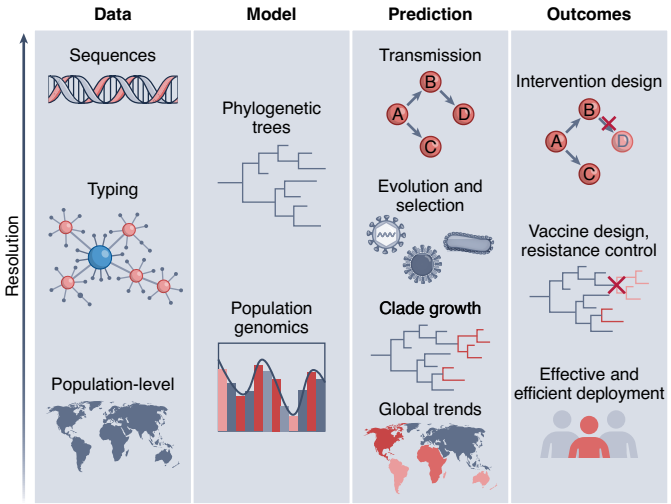

Fig. 1 | Data, models, predictions and outcomes may cover multiple levels of resolution. For example, data may comprise sequences, from which we wish to model sequence types or genotypes, to forecast global disease trends and thereby design an efficient resource deployment strategy. Alternatively, the composition of the pathogen population by larger sub-populations (for example, serotype or variant) may be the focal level for genetic diversity. There are myriad combinations, from which the scientist must determine the optimum scale at each step.

The potential of genomic data for forecasting in public health

Non-genomic mathematical models for infectious disease forecasting are widely applied in public health. There is a demand for models that allow public health agencies to prepare for expected demands on health services, vaccine stocks, mobilization of healthcare workers and communication campaigns[59]. Estimates of disease burden produced by the WHO[60] and others are used to compare the relative importance of different diseases and to determine where to allocate limited resources. Although progress was at first mostly limited to retrospective analyses using agent-based, compartmental or time-series models, over the past 20 years the implementation of epidemiological models for forecasting has become more commonplace. Three key factors driving progress have been the collection of high-resolution spatio-temporal data[61], the incorporation of more complex model features such as population structure and seasonal forcing, and computational advances in methods such as Markov chain Monte Carlo (MCMC)[62] and ABC[63].

We argue that the same steps towards forecasting should be taken with analyses incorporating genomic data. This need has previously been recognized[64], but with the collection of more longitudinal genomic data, there is increased opportunity. The application of genomic data to real-time analyses is currently limited, despite a marked increase in forecasting and 'nowcasting' analyses[65]. Although the COVID-19 pandemic has seen a plethora of research in forecasting[65,66] and genomic epidemiology[67,68], these analyses have remained largely separate. In other contexts, real-time sequencing of viruses is facilitating prediction from genomic data[69,70], although so far analyses have largely been descriptive.

The importance of forecasting in public health has been widely argued[71] and further emphasized during the COVID-19 pandemic with many public health organizations turning to mathematical models for regular jurisdictional forecasts, despite uncertainties. Incorporating genomic data into predictive models will offer new opportunities.

https://doi.org/10.1038/s41564-022-01233-6

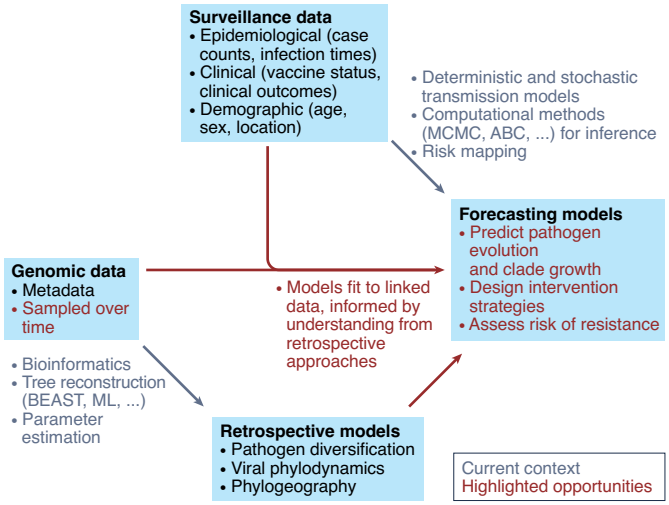

Fig. 2 | Depiction of current research directions and the opportunities highlighted in this work. We primarily focus on mechanistic models for forecasting, but these approaches can also include statistical or empirical models. Tree reconstruction methods include Bayesian Evolutionary Analysis Sampling Trees (BEAST) and maximum likelihood (ML).

For example, existing mathematical models for malaria incorporate population structures and immune selection to design drug resistance control strategies[72]. Genomics would seem a natural tool to extend this, although existing analyses have been more focused on genome description and identification of vaccine candidate antigens[73].

On the other hand, some findings in epidemiological models have been at odds with observations in genomics. Compartmental epidemiological models often predict competitive exclusion by the 'fittest' strains[59,74]. However, genomic studies have observed consistent strain diversity even in competing pathogen populations, such as long-term coexistence of drug-sensitive and drug-resistant strains of *S. pneumoniae*[75] and apparent frequency-dependent selection in *S. pneumoniae* and *E. coli*[76,77]. Further incorporation of genomics into epidemiological modelling could help to reconcile these contradictory perspectives and ensure that models can capture realistic diversity.

Early in the West African Ebola virus outbreak of 2013–2016, phylogenetic tools were used to trace the outbreak source and to characterize transmission patterns[78]. However, the majority of collected sequences could not be linked to individual case records[79], limiting applicability to modelling or forecasting, and studies that did include this were largely descriptive rather than predictive[80]. With pre-planned collection of genomic data during outbreaks and the goal of epidemiological analysis in mind, we could more fully incorporate the additional information that genomic data offer. A major difficulty with analyses of outbreaks is that the epidemic process is usually only partially observed: rarely do we know when individuals are infected or who infected whom. Genomic data can give us insight to these unobserved processes. However, this comes with ethical concerns, particularly around source attribution where this may have legal consequences or lead to stigmatization or social harm, as with HIV[81]. There is a growing base of ethical guidelines specifically concerning genomic research, but phylogenetic reconstruction studies must still make careful decisions around the costs and benefits of their findings[81]. The use of genomic data to reconstruct outbreaks also brings logistical challenges, for example, in rapid data collection and processing[82].

Genomic data also offer opportunities to improve vaccine design through increased understanding of pathogen diversity dynamics, as is underway for influenza[38]. Epidemiological models have been widely used, for example, in estimations of herd immunity thresholds[74] and to formulate vaccine development and deployment strategies[83].

In HIV research, anti-retroviral therapies have been analysed using compartmental epidemiological models[84]. However, this modelling has been focused at serotype and genotype levels, remaining somewhat separate from the field of phylodynamics despite seeking to answer similar questions. As non-genomic models for intervention strategies are not purely retrospective but also predictive, if we could integrate the rich information contained in now readily collected genomic data, we would be much better placed to make accurate forecasts incorporating evolutionary change.

Outlook

There are key challenges facing both surveillance specialists and modellers for progress in forecasting from genomic data. Longitudinally collected genomic data are critical to study how patterns of evolution and transmission are changing in time. Genomic and epidemiological data can be challenging to link, particularly when these are collected by different groups with different goals. Understandably, sharing genomic data and linking individual-level data (genomic, epidemiological, clinical) raise ethical and privacy questions, among many barriers to data sharing in public health[85]. All of the above will require dialogues between data collectors, data users and methods developers.

For modellers, finding the right level of abstraction is a challenge (Fig. 1). In most cases we do not wish to predict sequences, but rather the abundance or prevalence of different subgroups or types, or to understand selection and quantify the risk of emergence of new phenotypes and the impact of disease. Aims might include projecting the rate of spread of resistance, the emergence of new resistance or variants of concern, how strongly selection may favour new phenotypes, whether there are existing mutational profiles that could combine to confer advantages and so on. This will require finding an appropriate balance in the trade-off between simple and complex models, as well as useful summary statistics or descriptions of genomic data and the relationships between genomes. Methods that use genomes to infer population structure have been developed in recent years[86], but methods that characterize interactions across these structures have not yet really been explored.

Similarly, a challenge of existing phylogenetic and phylodynamic approaches, such as those using birth–death and coalescent models, is that they usually assume that genetic diversity is phenotypically neutral and are therefore not well suited to forecast resistance or antigenic evolution. There are modelling approaches that account for different growth or death rates in different lineages, including multi-strain epidemiological models, genome-scale negative frequency-dependent selection models[76,77], multi-type birth–death models[41], the binary-state speciation and extinction framework[87] and its extensions, and estimates of selection coefficients or fitness using genomic data[17,88]. These may group sequences into types or variants and proceed with an assumption of phenotypic neutrality within these types or focus on identifying mutations that confer an advantage. In our view, methods that directly incorporate selection strengthen links between genomic data and forecasting, even where their main focus is not making forward-time projections.

Although machine learning methods have become widely used tools in modern statistical analysis, their application to forecasting given genomic and epidemiological data is not straightforward. In addition to the drawbacks of difficult-to-interpret 'black box' methods[89], machine learning approaches have been shown to struggle with genome-wide association studies[90] despite a larger amount of training data than generally available in forecasting analyses. One key problem is that of hidden population structures in genomic data: complex interacting and evolving populations are likely to have complex dependence structures. These structures, if not accounted for in the mathematical model, can cause confounding[91]. Selection bias in which isolates are collected and which are included in sequencing studies is also a challenge, which must be accounted for or avoided. For example, prioritizing

Table 2 | Areas for further predictive applications of pathogen genomic data

Target of prediction	Purpose	Data and sampling requirements	Public health applications
Future patterns of lineage growth, dominance and coexistence	Infer competition and synergy among lineages and within and between hosts	Representative and longitudinal but not necessarily dense sampling	Measure impact of interventions targeting one lineage, for example, with antibiotic resistance
	Identify patterns of geographic spread	Representative and longitudinal but not necessarily dense sampling	Estimate likely impact of border measures; identify geographic sources and sinks
	Estimate selection coefficients and thereby predict population composition over time	Fractions of samples that are the different sub-populations, over time	Timescale to potential change of clinical impact, for example, time until particular variant of concern dominance in SARS-CoV-2
Phenotypic change	Changes in transmission, immune evasion, propensity to cause severe disease; clinical impact	Genomic data linked to phenotypes of interest (epidemiological or clinical data)	Changing symptoms and severity by SARS-CoV-2 variant
Resistance, acquired and inherited	Emergence and development of AMR	Resistant and sensitive 'background' sequences (through time); fitness estimates	Order and pace of emergence of resistance to antibiotics
Incidence and prevalence over time	Estimate parameters of a population-level model, then use that model to forecast prevalence and incidence	Genomic data to inform multi-strain models	Scenarios for future burden, syndemics, economic and wider impacts
Identify outbreaks and patterns of transmission in outbreak settings	Inference for outbreaks in less densely sampled settings or with more complex population structure	Linked epidemiological and genomic data, relatively densely sampled in the setting	Transmission timescales inform intervention timing; likely transmission contacts inform effective interventions
	Transmission prediction based on outbreak reconstruction, given phenotypic traits	Linked epidemiological and genomic data, relatively densely sampled in the setting	Identify risk factors for transmission; forecast outbreak sizes
Impacts of future public health intervention	Impacts of antibiotic usage on resistance	Data informing emergence and spread of resistance; validated models relating this to policy and usage data	Implications for antibiotic stewardship
	Impact of polyvalent vaccines on pathogen population composition	Longitudinal representative data, in the context of interventions	Design interventions to leave a more benign pathogen population, lower disease burden

We list some of the areas of inference, along with illustrative purposes, comments about data needs and example applications. This is by no means an exhaustive list, and descriptions are necessarily brief.

outbreaks for sequencing could lead to spurious conclusions of higher transmissibility if sampling strategy is not accounted for.

In both public health and evolution, we will require interpretable models that contain explanations of the predictions they make. This has ethical motivations in medical fields—patient safety, trust, concerns over unintended sociodemographic biases—and has been legally mandated, for example, by the European General Data Protection Regulation, which states that when personal data are used, the decision-making system of a model must be traceable and explainable[92]. This motivates the use of mechanistic and/or statistical models rather than, or in combination with, machine learning models and ties in with identified challenges in explainable AI[93]. The dimension of time, not present in genome-wide association studies or the vast majority of existing genomics applications in mathematics and statistics, also introduces complexities. We now require understanding of the interactions between organisms at different scales, including competition, horizontal gene transfer, synergy and niche differentiation. Over long periods where the drifting dynamics of these interactions may not yet be well understood, forecasting may not be feasible. However, mechanistic approaches offer the opportunity to model these behaviours at different levels and allow interpretation of the interactions between them.

The 'curse of dimensionality' is particularly challenging with genomic data—for example, the number of potential genotypes increases exponentially with the number of loci considered. It is not possible to sample every combination, resulting in limitations to the feasibility of prediction, and increasing the potential for population stratification confounding and the computational complexity. Factors affecting the success of different pathogen genotypes may depend on complex interactions between large numbers of loci and numerous environmental factors[94]. The dependence structures may not be known in advance, and the set of possible dependencies is large. Computational techniques will therefore need to infer or otherwise account for unknown dependence structures and overcome the dimensionality problems they introduce.

Throughout this Perspective, we have discussed areas of research that would benefit from incorporation of genomic data for forecasting. From existing forecasting approaches in infectious disease modelling to existing approaches in phylogenetics and other genomic research that have had limitations when it comes to prediction, further combining approaches and developing new methods at the intersection will unlock new possibilities, particularly with the ever-increasing availability of longitudinally sampled pathogen genomic data (Fig. 2). Many research efforts are already moving in this direction, but we take a more speculative view in Table 2 of where future research could focus, for what purpose, what the data and sampling challenges will be and where this may be most applicable to public health.

Conclusions

We propose that there are high potential benefits to developing forecasting methods that can combine genomic data with epidemiological, clinical and surveillance system data. This will require combining existing techniques in novel ways (Fig. 2 and Table 2) and developing new approaches. If we can incorporate pathogen dynamics and evolution into existing forecasting approaches, there is scope to make more robust predictions. Similarly, methods that fit models to genomic data to estimate epidemiological parameters can be extended for forecasting by incorporating knowledge of the underlying generative processes. Although the application of machine learning methods to

genomic epidemiological analyses has limitations, there is scope to further integrate them. Linking mechanistic models to machine learning approaches can help to motivate their structure, interpret their outputs or gain intuition about the mechanistic behaviours behind forecasts. All of the above has been made possible by tremendous efforts to collect and compile genomic data into publicly available repositories and would be further facilitated by (1) more longitudinally collected and representative sequences and (2) linkage to epidemiological, demographic and clinical information where feasible. Over the past 20 years, the rich information that genomic data contain has been successfully applied to retrospective epidemiological analyses. The next step is for genomic data to help us understand more about possible futures to come.

References

1. Land, M. et al. Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics* **15**, 141–161 (2015).
2. Kennedy, D. & Read, A. Why does drug resistance readily evolve but vaccine resistance does not? *Proc. R. Soc. B* **284**, 20162562 (2017).
3. Volberding, P. A. & Deeks, S. G. Antiretroviral therapy and management of HIV infection. *Lancet* **376**, 49–62 (2010).
4. Petrova, V. N. & Russell, C. A. The evolution of seasonal influenza viruses. *Nat. Rev. Microbiol.* **16**, 47–60 (2018).
5. Dyson, L. et al. Possible future waves of SARS-CoV-2 infection generated by variants of concern with a range of characteristics. *Nat. Commun.* **12**, 5730 (2021).
6. *Multidrug and Extensively Drug-Resistant TB (M/XDR-TB): 2010 Global Report on Surveillance and Response* (World Health Organization, 2010).
7. Wellems, T. & Plowe, C. Chloroquine-resistant malaria. *J. Infect. Dis.* **184**, 770–776 (2001).
8. Chewapreecha, C. et al. Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet.* **46**, 305–309 (2014).
9. Croucher, N. J. et al. Rapid pneumococcal evolution in response to clinical interventions. *Science* **331**, 430–434 (2011).
10. Doyle, R. M. Direct whole-genome sequencing of sputum accurately identifies drug-resistant Mycobacterium tuberculosis faster than MGIT culture sequencing. *J. Clin. Microbiol.* **56**, e00666-18 (2018).
11. Feuerriegel, S. et al. PhyResSE: a web tool delineating *Mycobacterium tuberculosis* antibiotic resistance and lineage from whole-genome sequencing data. *J. Clin. Microbiol.* **53**, 1908–1914 (2015).
12. Domman, D. et al. Defining endemic cholera at three levels of spatiotemporal resolution within Bangladesh. *Nat. Genet.* **50**, 951–955 (2018).
13. Williamson, D. A. et al. Bridging of Neisseria gonorrhoeae lineages across sexual networks in the HIV pre-exposure prophylaxis era. *Nat. Commun.* **10**, 3988 (2019).
14. Wensing, A. M. et al. 2019 update of the drug resistance mutations in HIV-1. *Top. Antivir. Med.* **27**, 111 (2019).
15. Simmonds, P. Genetic diversity and evolution of hepatitis C virus–15 years on. *J. Gen. Virol.* **85**, 3173–3188 (2004).
16. Lemey, P., Rambaut, A. & Pybus, O. G. HIV evolutionary dynamics within and among hosts. *Aids Rev.* **8**, 125–140 (2006).
17. Obermeyer, F. et al. Analysis of 6.4 million SARS-CoV-2 genomes identifies mutations associated with fitness. *Science* **376**, 1327–1332 (2022).
18. Grenfell, B. et al. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science* **303**, 327–332 (2004).
19. Volz, E. M., Kosakovsky Pond, S. L., Ward, M. J., Leigh Brown, A. J. & Frost, S. D. Phylodynamics of infectious disease epidemics. *Genetics* **183**, 1421–1430 (2009).
20. Kühnert, D., Stadler, T., Vaughan, T. G. & Drummond, A. J. Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death SIR model. *J. R. Soc. Interface* **11**, 20131106 (2014).
21. Boskova, V., Bonhoeffer, S. & Stadler, T. Inference of epidemiological dynamics based on simulated phylogenies using birth–death and coalescent models. *PLoS Comput. Biol.* **10**, e1003913 (2014).
22. Volz, E. et al. Phylodynamic analysis to inform prevention efforts in mixed HIV epidemics. *Virus Evol.* **3**, vex014 (2017).
23. O'Neill, J. Antimicrobial Resistance: Tackling a Crisis for the Health and Wealth of Nations. *Review on Antimicrobial Resistance* (2014).
24. Colijn, C. et al. What is the mechanism for persistent coexistence of drug-susceptible and drug-resistant strains of *Streptococcus pneumoniae*? *J. R. Soc. Interface* **7**, 905–919 (2010).
25. Knight, G. M. et al. Mathematical modelling for antibiotic resistance control policy: do we know enough? *BMC Infect. Dis.* **19**, 1011 (2019).
26. Mukherjee, S. et al. Genomes OnLine Database (GOLD) v. 8: overview and updates. *Nucleic Acids Res.* **49**, D723–D733 (2021).
27. Brister, J. R., Ako-Adjei, D., Bao, Y. & Blinkova, O. NCBI viral genomes resource. *Nucleic Acids Res.* **43**, D571–D577 (2015).
28. Shu, Y. & McCauley, J. GISAID: global initiative on sharing all influenza data—from vision to reality. *Eurosurveillance* **22**, 30494 (2017).
29. Davis, J. J. et al. The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic Acids Res.* **48**, D606–D612 (2020).
30. Muir, P. et al. The real cost of sequencing: scaling computation to keep pace with data generation. *Genome Biol.* **17**, 53 (2016).
31. McArthur, A. et al. The comprehensive antibiotic resistance database. *Antimicrob. Agents Chemother.* **57**, 3348–3357 (2013).
32. Lakin, S. et al. MEGARes: an antimicrobial resistance database for high throughput sequencing. *Nucleic Acids Res.* **45**, D574–D580 (2017).
33. Arango-Argoty, G. et al. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* **6**, 23 (2018).
34. UniProt Consortium UniProt: a hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015).
35. Lewis, F., Hughes, G. J., Rambaut, A., Pozniak, A. & Brown, A. J. L. Episodic sexual transmission of HIV revealed by molecular phylodynamics. *PLoS Med.* **5**, e50 (2008).
36. Faria, N. R. et al. The early spread and epidemic ignition of HIV-1 in human populations. *Science* **346**, 56–61 (2014).
37. Volz, E. M. & Didelot, X. Modeling the growth and decline of pathogen effective population size provides insight into epidemic dynamics and drivers of antimicrobial resistance. *Syst. Biol.* **67**, 719–728 (2018).
38. Łuksza, M. & Lässig, M. A predictive fitness model for influenza. *Nature* **507**, 57–61 (2014).
39. Stadler, T. et al. Estimating the basic reproductive number from viral sequence data. *Mol. Biol. Evol.* **29**, 347–357 (2011).
40. Geidelberg, L. et al. Genomic epidemiology of a densely sampled COVID-19 outbreak in China. *Virus Evol.* **7**, veaa102 (2021).
41. Stadler, T. & Bonhoeffer, S. Uncovering epidemiological dynamics in heterogeneous host populations using phylogenetic methods. *Phil. Trans. R. Soc. B* **368**, 20120198 (2013).
42. Rasmussen, D. A. & Stadler, T. Coupling adaptive molecular evolution to phylodynamics using fitness-dependent birth–death models. *eLife* **8**, e45562 (2019).
43. Lemey, P., Rambaut, A., Drummond, A. & Suchard, M. Bayesian phylogeography finds its roots. *PLoS Comput. Biol.* **5**, 1–16 (2009).

44. Ratmann, O., Donker, G., Meijer, A., Fraser, C. & Koelle, K. Phylodynamic inference and model assessment with approximate Bayesian computation: influenza as a case study. *PLoS Comput. Biol.* **8**, 12 e1002835 (2012).
45. Lewitus, E. & Morlon, H. Characterizing and comparing phylogenies from their Laplacian spectrum. *Syst. Biol.* **65**, 495–507 (2015).
46. Liu, P., Biller, P., Gould, M. & Colijn, C. Analyzing phylogenetic trees with a tree lattice coordinate system and a graph polynomial. *Syst. Biol.* https://doi.org/10.1093/sysbio/syac008 (2022).
47. Kim, J., Rosenberg, N. A. & Palacios, J. A. Distance metrics for ranked evolutionary trees. *Proc. Natl Acad. Sci. USA* **117**, 28876–28886 (2020).
48. Hall, M., Woolhouse, M. & Rambaut, A. Epidemic reconstruction in a phylogenetics framework: transmission trees as partitions of the node set. *PLoS Comput. Biol.* **11**, e1004613 (2016).
49. Didelot, X., Fraser, C., Gardy, J. & Colijn, C. Genomic infectious disease epidemiology in partially sampled and ongoing outbreaks. *Mol. Biol. Evol.* **34**, 997–1007 (2017).
50. McCloskey, R., Liang, R. & Poon, A. Reconstructing contact network parameters from viral phylogenies. *Virus Evol.* **2**, vew029 (2016).
51. McLaughlin, A. et al. Concordance of HIV transmission risk factors elucidated using viral diversification rate and phylogenetic clustering. *Evol. Med. Public Health* **9**, 338–348 (2021).
52. Rasmussen, D. A., Ratmann, O. & Koelle, K. Inference for nonlinear epidemiological models using genealogies and time series. *PLoS Comput. Biol.* **7**, e1002136 (2011).
53. Cottam, E. et al. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc. Biol. Sci.* **275**, 887–895 (2008).
54. Jombart, T. et al. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Comput. Biol.* **10**, e1003457 (2014).
55. Bedford, T., Cobey, S., Beerli, P. & Pascual, M. Global migration dynamics underlie evolution and persistence of human influenza A (H3N2). *PLoS Pathog.* **6**, e1000918 (2010).
56. Maurano, M. T. et al. Sequencing identifies multiple early introductions of SARS-CoV-2 to the New York City region. *Genome Res.* **30**, 1781–1788 (2020).
57. Miller, D. et al. Full genome viral sequences inform patterns of SARS-CoV-2 spread into and within Israel. *Nat. Commun.* **11**, 5518 (2020).
58. Colijn, C. et al. The need for linked genomic surveillance of SARS-CoV-2. *Can. Commun. Dis. Rep* **48**, 131–139 (2022).
59. Knight, G. et al. Bridging the gap between evidence and policy for infectious diseases: how models can aid public health decision-making. *Int. J. Infect. Dis.* **42**, 17–23 (2016).
60. Mathers, C. D. History of global burden of disease assessment at the World Health Organization. *Arch. Public Health* **78**, 77 (2020).
61. Lowe, R. et al. Spatio-temporal modelling of climate-sensitive disease risk: towards an early warning system for dengue in Brazil. *Comput. Geosci.* **37**, 371–381 (2011).
62. O'Neill, P. A tutorial introduction to Bayesian inference for stochastic epidemic models using Markov chain Monte Carlo methods. *Math. Biosci.* **180**, 103–114 (2002).
63. McKinley, T., Ross, J., Deardon, R. & Cook, A. Simulation-based Bayesian inference for epidemic models. *Comput. Stat. Data Anal.* **71**, 434–447 (2014).
64. Gandon, S., Day, T., Metcalf, C. J. E. & Grenfell, B. T. Forecasting epidemiological and evolutionary dynamics of infectious diseases. *Trends Ecol. Evol.* **31**, 776–788 (2016).
65. Wu, J. T., Leung, K. & Leung, G. M. Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak originating in Wuhan, China: a modelling study. *Lancet* **395**, 689–697 (2020).
66. Anastassopoulou, C., Russo, L., Tsakris, A. & Siettos, C. Data-based analysis, modelling and forecasting of the COVID-19 outbreak. *PLoS ONE* **15**, e0230405 (2020).
67. Lu, J. et al. Genomic epidemiology of SARS-CoV-2 in Guangdong Province, China. *Cell* **181**, 997–1003 (2020).
68. Stockdale, J. E. et al. Genomic epidemiology offers high resolution estimates of serial intervals for COVID-19. Preprint at *medRxiv* https://doi.org/10.1101/2022.02.23.22271355 (2022).
69. Siddle, K. et al. Genomic analysis of Lassa virus from the 2018 surge in Nigeria. *N. Engl. J. Med.* **379**, 1745–1753 (2018).
70. Quick, J. et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature* **47**, 309–330 (2016).
71. Myers, M., Rogers, D., Cox, J., Flahault, A. & Hay, S. Forecasting disease risk for increased epidemic preparedness in public health. *Adv. Parasitol.* **47**, 309–330 (2000).
72. Mandal, S., Sarkar, R. & Sinha, S. Mathematical models of malaria—a review. *Malar. J.* **10**, 202 (2011).
73. Le Roch, K. G., Chung, D.-W. & Ponts, N. Genomics and integrated systems biology in *Plasmodium falciparum*: a path to malaria control and eradication. *Parasite Immunol.* **34**, 50–60 (2012).
74. Anderson, R. & May, R. *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ. Press, 1991).
75. Lipsitch, M., Colijn, C., Cohen, T., Hanage, W. & Fraser, C. No coexistence for free: neutral null models for multistrain pathogens. *Epidemics* **1**, 2–13 (2009).
76. Corander, J. et al. Frequency-dependent selection in vaccine-associated pneumococcal population dynamics. *Nat. Ecol. Evol.* **1**, 1950–1960 (2017).
77. McNally, A. et al. Diversification of colonization factors in a multidrug-resistant *Escherichia coli* lineage evolving under negative frequency-dependent selection. *MBio* **10**, e00644-19 (2019).
78. Gire, S. et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science* **345**, 1369–1372 (2014).
79. Cori, A. et al. Key data for outbreak evaluation: building on the Ebola experience. *Phil. Trans. R. Soc. B* **372**, 1369–1372 (2017).
80. Baize, S. et al. Emergence of Zaire Ebola virus disease in Guinea. *N. Engl. J. Med.* **371**, 1418–1425 (2014).
81. Coltart, C. E. et al. Ethical considerations in global HIV phylogenetic research. *Lancet HIV* **5**, e656–e666 (2018).
82. Reich, N. et al. Challenges in real-time prediction of infectious disease: a case study of dengue in Thailand. *PLoS Negl. Trop. Dis* **10**, e0004761 (2016).
83. Keeling, M., Woodhouse, M., May, R., Davies, G. & Grenfell, B. Modelling vaccination strategies against foot-and-mouth disease. *Nature* **421**, 136–142 (2003).
84. Smith, D. & Mideo, N. Modelling the evolution of HIV-1 virulence in response to imperfect therapy and prophylaxis. *Evol. Appl.* **10**, 297–309 (2017).
85. van Panhuis, W. G. et al. A systematic review of barriers to data sharing in public health. *BMC Public Health* **14**, 1144 (2014).
86. Bryc, K. et al. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc. Natl Acad. Sci. USA* **107**, 786–791 (2010).
87. Maddison, W. P., Midford, P. E. & Otto, S. P. Estimating a binary character's effect on speciation and extinction. *Syst. Biol.* **56**, 701–710 (2007).
88. Jankowiak, M., Obermeyer, F. H. & Lemieux, J. E. Inferring selection effects in SARS-CoV-2 with Bayesian viral allele selection. Preprint at *bioRxiv* https://doi.org/10.1101/2022.05.07.490748 (2022).
89. Xu, C. & Jackson, S. Machine learning and complex biological data. *Genome Biol.* **20**, 76 (2019).
90. Kim, Y. et al. Evaluation of random forests performance for genome-wide association studies in the presence of interaction effects. *BMC Proc.* **3**, S64 (2009).

91. Sul, J., Martin, L. & Eskin, E. Population structure in genetic studies: confounding factors and mixed models. *PLoS Genet.* **14**, e1007309 (2018).
92. Mourby, M., Cathaoir, K. Ó. & Collin, C. B. Transparency of machine-learning in healthcare: the GDPR & European health law. *Comput. Law Security Rev.* **43**, 105611 (2021).
93. Goebel, R. et al. Explainable AI: the new 42? In *2nd International Cross-Domain Conference for Machine Learning and Knowledge Extraction (CD-MAKE)* hal-01934928 295–303 (IFIP, 2018).
94. Thomas, M. S. & Wigneshweraraj, S. Regulation of virulence gene expression. *Virulence* **5**, 832–834 (2014).
95. Foley, B. et al. *Los Alamos HIV Sequence Compendium 2018* (Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, 2018).

Acknowledgements

This work was supported by the grant of the Federal Government of Canada's Canada 150 Research Chair programme to C.C.

Author contributions

J.E.S., P.L. and C.C. all contributed to the conceptualization, writing and revision of this manuscript. All authors approve the submitted version and agree to be personally accountable for the work.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence should be addressed to Caroline Colijn.

Peer review information *Nature Microbiology* thanks Katia Koelle and Art Poon for their contribution to the peer review of this work.