The opportunity of data-driven services for viral genomic surveillance

Anna Bernasconi  
Dipartimento di Elettronica, Informazione e Bioingegneria  
Politecnico di Milano  
Milan, Italy  
anna.bernasconi@polimi.it

Abstract—The recent COVID-19 pandemic has posed novel challenges to the big data and knowledge management community. The unprecedented availability of viral genomes on public databases has made possible the data-driven exploration of viruses’ evolution (especially of SARS-CoV-2, the virus responsible for the disease). Properties of data and knowledge in the genomic and virological domain may fuel data science methods for the identification and possible prediction of critical phenomena, such as the emergence of variants with improved transmissibility/virulence and recombined strains. A number of tools have been produced to explore the variants’ trends or suggest hypotheses on the evolutionary mechanisms of the virus. In this perspective, we elaborate on plausible directions of this field of research, which are still applicable to the SARS-CoV-2 virus but may become even more relevant in the context of new outbreaks (e.g., monkeypox, malaria, diphtheria). Expressly, we point to 1) data-driven identification of mutations or variants with potential impact; 2) data-driven identification of recombination events – creating opportunities to overcome selective pressure and adapt to new environments and hosts (e.g., livestock or humans). These directions can be framed within genomic surveillance measures, characterized by the possibility of tracking viruses by using their genome, which is collected, sequenced, and submitted to public databases by laboratories around the world. If successful, genomic surveillance substantially supports the understanding of novel viral pathogens and of their dangerousness in terms of prevalence, infectivity, and transmissibility; the implemented services can be of great utility to decision-makers in healthcare. Here, we draw current trends, challenges, and future directions of data-driven services for genomic surveillance.

Index Terms—Genomic surveillance, big data analytics, big data services, virology, pathogen evolution

I. INTRODUCTION

The developments in high throughput DNA sequencing technologies (called Next Generation Sequencing [1]) and the recent reduction in costs [2] represent an important achievement and turning point for modern molecular biology. NGS technologies have become routinely employed in different fields of genomic research. Different sequencing platforms and sample preparation approaches, in laboratories worldwide, contributed to a revolution in the detection and discovery of viral genomes. The spread and evolution of pathogens can be currently tracked by the comparison of NGS-based genomic sequences [3] derived from different specimens and collected at different locations, a methodology called genomic surveillance. Effective tools for pathogen genomics surveil-

lance represent the first defense lines against current and future epidemics. Even if effective surveillance does not require the sequencing of a specimen from every case, it is important that enough sequence data is collected from representative populations to detect new variants and monitor trends in circulating variants. Genomic surveillance for SARS-CoV-2 has thus far been based on the wide availability of genomes deposited around the world [4] and has been considered of the utmost importance by the World Health Organization – which has also established a working group dedicated to it [5].

Infectious diseases that emerged in the last decade (see, e.g., SARS, MERS, Zika, and Ebola) and more recently (SARS-CoV-2), have demonstrated the importance of genomic data to study the evolution of pathogens, tracking their spread, and generating a deeper understanding of infectious diseases. Even when they are not discussed in the news, outbreaks are continuously detected; indeed, the US Centers for Disease Control and Prevention (CDC, [6]) reports 13 U.S.-Based Outbreaks (including, e.g., fungal meningitis, Salmonella infections, hepatitis A), seven Travel Notices Affecting International Travelers (including, e.g., Fungal Infections in Mexico, Malaria in Costa Rica, or Nipah Virus in Bangladesh), and four International Outbreaks, i.e., the Coronavirus Disease 2019 – announced January 2020, the U.S. monkeypox outbreak – announced May 2022, and two Ebola outbreaks in August 2022 respectively in the Democratic Republic of the Congo and Uganda.

All viruses change when they replicate and spread in a population. Most of the acquired mutations in their genome are silent (i.e., not translated into proteins) or neutral (i.e., unlikely to change key features), meaning that they do not affect the virus’ ability to spread or escape the immune response because they do not alter the major proteins involved in infection and transmission. Although rare, mutations that instead improve these abilities can eventually be selected, leading to a competitive advantage over the wild (i.e., original) type of virus. When this happens, novel forms of the pathogen with improved epidemiological features can emerge and circulate in a population—see “variant of interest” (VOI) or “variant of concern” (VOC) as named by the World Health Organization (WHO) [7].

Because of the complexity of viral infectious diseases,
understanding the origin, the evolution of the virus, and its molecular basis requires continuously updated data, allowing to produce investigations that rely on it and can consequently shape timely effective responses. During COVID-19 times, the bioinformatics community has observed the production of an exorbitant amount of data [8]; the total number of collected and sequenced genomes of SARS-CoV-2 available worldwide went from a few hundred in March 2020, up to about one hundred thousand in August 2020, and to more than 10 million in 2022, reaching 15.6 million at the time of writing, in June 2023 (about 3.5 years since the first outbreak). There has been a continuous deposition of SARS-CoV-2 sequences to public repositories; a similar approach has been adopted for Influenza and Monkeypox data [9]. All other viruses have reference databases, although smaller amounts of data [10].

In this paper, we focus on two threads of viral genomic surveillance research:

- data-driven variant identification;
- data-driven recombination events identification.

In these fields, several methodologies are being studied – some of which have been and can still be offered in the form of data-driven services, made available to the research community or to the wider public. In the following, Section II presents the ingredients of the domain that are useful to understand the remaining of the paper; Section III describes the research field in charge of monitoring the rise of viral variants, reporting a number of tools usable by a wide public for studying variants; Section IV reports on the past and current efforts for the identification of viral recombination; finally, Section V describes the open challenges in the genomic surveillance context and Section VI concludes the chapter with future directions.

II. BACKGROUND

Until the COVID-19 pandemic, the data and knowledge referring to viral domains have been the prerogative of molecular biologists and virologists. Starting then, the unprecedented richness of information has attracted the attention of a much broader research community, with the consequent rise of organization and systematization efforts.

In [11] we presented CoV2K, a high-level description of information related to SARS-CoV-2. Here, we denote as Knowledge the established information about the virus and as Data the actual genomes or their fragments, which are continuously produced and represent data points for analysis. The knowledge ecosystem includes information on variants (with their names, presenting organization, and computational methods to determine them); their effects reported in research studies or alternative evidence (such as their resistance to monoclonal antibodies, convalescent/vaccine sera, transmissibility, or virulence); and their characterization (in terms of sets of mutations, with specific positions and molecular changes, along the structure of the viral genome or specific proteins). Moreover, it includes the characteristics of mutations due to their original and alternative nucleotide or amino acid residues – these present several features, such as changes in polarity, hydrophobicity, or charge. Finally, it includes the definition of particular regions of the genome with given functions.

The data ecosystem, instead, includes information about actual genomes – previously described in the Viral Conceptual Model (VCM, [12]). These have been collected and sequenced, involving the use of specific sequencing platforms (with an associated accuracy and coverage) and algorithms (e.g., to perform genome assembly and variant calling). Additionally, it includes epitopes, which are strings of amino acid residues from a virus protein that can be recognized by antibodies or other host receptors.

In the CoV2K model, knowledge and data are connected by many relationships, driving a process of data and knowledge integration. Next, we delve into the details of the two core concepts of the model, i.e., mutations and variants.

Mutations are found either at the DNA or RNA level (depending on the kind of virus – SARS-CoV-2 is RNA-based), called nucleotide mutations, or at the protein level, called amino acid changes. Nucleotide mutations occur at specific positions of the virus genome, causing deletions, insertions, or substitutions. They have a position, where the reference nucleotide is changed into an alternative one, affecting a certain length of the sequence. As an example, the notation A23403G indicates that the 23403rd nucleotide of the sequence, which was Adenine in the virus wild-type, has been changed into a Guanine. These mutations can be silent (neutral or synonymous) when they do not change the amino acid sequence, or non-synonymous, when they change the translated amino acid sequence. They also have a position, which is relevant to understand which protein functionalities may be impacted. The most common notation for amino acid changes is a string that mentions the protein and the mutation signature. For example, S:D614G denotes a substitution at the 614th position of the Spike protein, from the wild-type Aspartic Acid (D) to Glycine (G). Since several mutations may jointly produce stronger effects, the co-occurrence of nucleotide mutations or amino acid changes on a single viral genome is also relevant.

Variants are forms of a virus that are considerably different from its original wild-type, as they accumulated a set of amino acid changes that characterize their phenotypic characteristics [13]. The definition of variants is produced with a phylogenetic analysis based on the use of phylogenetic trees [14]; these trees describe the precise chain of evolutionary changes leading from one viral genome to the next. As a result, a separation of viral sequences into lineages is defined – these share common ancestries and the same amino acid changes. Based on phylogenetics, wide complex taxonomies are produced to categorize viruses, continuously challenged by new strains [15].

In the case of SARS-CoV-2, a number of nomenclatures are used to denote specific lineages. The most common ones are those proposed by Pangolin [16], Nextstrain [17], and GISAID [9]. Given the societal and mass-media impact of the pandemic, the WHO proposed a naming scheme for variants aiming for wide adoption, based on the Greek alphabet [18].
During the years of the COVID-19 pandemic, several variants have attracted the attention of the research community and the general public. The most notable ones are Alpha, Delta, and Omicron – in four forms (respectively denoted as Omicron 1, 2, 4, and 5). Previously, other variants of interest were indicated, named Beta, Gamma, Epsilon, Zeta, Eta, Theta, Iota, Kappa, Lambda, and Mu. At the time of writing (June 2023), the WHO does not recognize any circulating variant of concern but only two circulating variants of interest, i.e., XBB.1.5 and XBB.1.16, respectively named Kraken and Arcturus, according to the unofficial nomenclature proposed on Twitter by T. Ryan Gregory [19]. These derive from recombination events of two sub-lineages of Omicron 2 and have recently threatened vaccine efficacy [20], [21]. Aside from WHO, surveillance of variants and their effects is also performed by national and international organizations such as the US Centers for Disease Control and Prevention, the European Center for Disease and Control, and Public Health England.

III. DATA-DRIVEN SARS-COV-2 VARIANT IDENTIFICATION

A. Datasets

The landscape of relevant resources and initiatives dedicated to data collection and retrieval of virus sequences was surveyed in [22]. The space of contributors can be partitioned by considering general institutions that host data sequences, primary sequence deposition databases, and tools provided for directly querying and searching them. The three main organizations providing open-source viral sequences are NCBI (US), EMBL-EBI (Europe), and DDBJ (Japan); they operate within the broader context of the International Nucleotide Sequence Database Collaboration [23]. NCBI hosts the two, so far, most relevant open viral sequence databases: GenBank [10] contains the annotated collection of all publicly available DNA and RNA sequences; RefSeq [24] provides a stable reference for genome annotation, gene identification and characterization, and mutation or polymorphism analysis. GenBank has been continuously updated thanks to the abundant sharing of multiple laboratories and data contributors around the world (SARS-CoV-2 nucleotide sequences have increased from about 300 around the end of March 2020 to 7.1 million as of June 2023). EMBL-EBI hosts the European Nucleotide Archive [25], which also accepts submissions of raw sequencing data, sequence assembly information, and functional annotations. While the INSDC consortium provides full open access to sequences, the GISAID initiative [9] was created in 2008 with the explicit purpose of offering an alternative to traditional public-domain data archives, as many scientists hesitated to share influenza data. GISAID hosts EpiFlu™, a large sequence database, which started its mission for influenza data and has expanded with EpiCoV™ having a particular focus on the SARS-CoV-2 pandemic (about 15.6 million sequences as of June 2023). In 2022, the initiative extended its interest also to Monkeypox genomes. Finally, COG-UK [26] is a national-based initiative launched in March 2020 thanks to big financial support from three institutional UK partners; alone, it contributed 20% of the world production of SARS-CoV-2 genome sequences.

B. Methods

The emergence of mutation and variants has traditionally been studied with phylogenetics methods, which accumulate individual sequences along the tree of the virus, built incrementally. When a branch becomes rich in sequences, it is a candidate variant to be investigated. The phylogenetic analysis takes into account the entire history of the viral genome evolution. However, the wide abundance of data occurred for SARS-CoV-2 has motivated the experimentation of other data mining techniques, i.e., methods that study the virus characteristics independently from the traditional phylogenetic techniques.

Methods initially concentrated on the monitoring of individual amino acid changes, such as the rise of the Spike protein D614G mutation [27], which soon became prevalent worldwide or the spread of the Spike A222V mutation that was supposedly generated in Spain and then spread throughout Europe during the Summer of 2020 [28].

As the COVID-19 pandemic progressed, research interests shifted toward the study of groups mutations, co-occurring on the same genomes – possibly leading to increased transmission rates and changed antigenicity of the SARS-CoV-2 virus, hampering testing, treatment, and vaccine development [29]–[31].

Considerable efforts have been dedicated to building surveillance systems that take advantage of this big data corpus, by employing temporal analysis of mutations to assist in the identification of candidate variants and their possible effects [32]. A number of studies have described typical SARS-CoV-2 mutational profiles across different countries and regions [33], proposing statistical indicators for location-based mutation evolution [34] and observing changes that become recurrently prevalent in different locations, thus suggesting selective advantages [35]. Several methods have been developed to study prevalent SARS-CoV-2 mutations over time and how they behave in a coordinated manner. Basically, many works consider functions that describe the prevalence of different mutations and – when a number of these are behaving similarly – they recognize a possible distinct variant. In Table I we report the salient characteristics of some works, including those based on phenetic clustering of prevalent SARS-CoV-2 mutations over time [32]–[34]; time-series clustering [35]–[37]; and weighted networks of frequency trajectories of mutations [38], [39].

We here selected only works that study these phenomena on a pandemic scale, whereas we exclude the vast literature analyzing the evolution of mutational patterns that are typical of a specific geographical area.

C. Services

Several tools arose during the pandemic, providing services that leverage the big datasets of SARS-CoV-2 sequences of international institutions such as GISAID and NCBI Virus (GenBank). Specifically, in the first two years, many online
With time passing, tools that are more sophisticated in their mutation and variant hunting purposes, have been offered to the community. Some form of a watch of variations is put in place by the World Health Organization – with several tools available to the public including the COVID-19 dashboard [31] – and GISAID, which integrates on its platform CoVizu [32] and CoVsurver [33]. Nextstrain [34] has a dedicated dashboard available to the public including the COVID-19 dashboard [35] – and GISAID, which integrates on its platform CoVizu [36] and CoVsurver [37]. Table I shows a selection of the tools with salient characteristics and differences.

### TABLE I

**OVERVIEW OF DATA-DRIVEN METHODS FOR VIRAL VARIANT IDENTIFICATION**

<table>
<thead>
<tr>
<th>Proponent</th>
<th>Study description</th>
</tr>
</thead>
</table>
| Yang et al. 2020     | **Methods:** Various clustering analyses.  
**Results:** Identified six types of strains and underlying signature single-mutations.  
**Highlight:** Suggested that single-mutations could become an important consideration in SARS-CoV-2 classification and surveillance. |
| Chiara et al. 2021   | **Methods:** Phylogenetic and phylodynamic methods are essential to study the spread and evolution of viruses; they are based on the assumption that the shared history of pathogens - when isolated |
from different hosts - can be described by a phylogenetic tree. Nonetheless, recombination does not follow this assumption, making it problematic to use phylogenetic methods to study recombining pathogens [56]. Thus, inference and computational big data-driven methods become necessary [57].

Recombination has not been highly prevalent in the three years of the SARS-CoV-2 pandemic and identification of early recombinant genomes has been difficult also due to the fact that the phylogenetic structure of SARS-CoV-2 is driven by a limited number of mutations (possibly often clustered in short regions of the genome). The contribution of recombination to the evolution of this virus has been considered generally low, although recently accelerating [58].

Many have studied the first recombination events and tried to delineate the extent to which recombination was expected to shape SARS-CoV-2 in the following years [59]–[61]. The first relevant recombination event has been reported to be between Alpha and Delta SARS-CoV-2 variants (Japan, second half of 2021) [62]. Delta and Omicron 1 co-circulated from November 2021 until February 2022: cases have been reported of co-infection [63]; one case appeared to be due to laboratory artifacts [64]. All these recombinants were soon out-competed by Omicron 2. A full report is provided by Focosi et al. [54]. Notably, the only Variants Of Interest by the WHO at the time of writing [7], i.e., XBB.1.5 and XBB.1.16, both derived from the lineage XBB, which recombined from two different descendants of Omicron 2.

Before SARS-CoV-2, viral recombinants were identified using algorithms implemented in 3SEQ [65] and RDP3 (Recombination Detection Program version 3) now evolved into RDP5 [66]. These algorithms aggregate several phylogenetics-based methods to test recombination hotspots and pinpoint patterns of interest with matrix-based visualizations. A number of studies applied these methods also to SARS-CoV-2 datasets [58], [60].

Table II presents the currently available recombination-detection methods (and connected services) that have been tested on SARS-CoV-2.

V. OPEN CHALLENGES

Genomic surveillance is the process of constantly monitoring pathogens and analyzing their genetic similarities and differences; it has become of worldwide interest since the first COVID-19 outbreak at the end of 2019, but can be put in place for any kind of pathogen. In these last three years, the fast emergence of SARS-CoV-2 mutations and their possibly severe epidemiological/immunological implications have called for continuous and worldwide monitoring of viral genomes.

Table II

<table>
<thead>
<tr>
<th>Service</th>
<th>Used data</th>
<th>Mutation search</th>
<th>Metadata search</th>
<th>Claimed purpose</th>
</tr>
</thead>
<tbody>
<tr>
<td>CoV-Spectrum</td>
<td>GISAID + Swiss</td>
<td>nucleotide; amino acid</td>
<td>date; location</td>
<td>Tracking of known variants; identification of new SARS-CoV-2 variants of concern.</td>
</tr>
<tr>
<td>COVID-19 CG</td>
<td>GISAID</td>
<td>amino acid</td>
<td>date; location</td>
<td>Tracking SARS-CoV-2 single-nucleotide mutations and lineages, useful to study SARS-CoV-2 transmission, evolution, diagnostics, therapeutics, vaccines, and intervention tracking.</td>
</tr>
<tr>
<td>outbreak.info</td>
<td>GISAID</td>
<td>lineage; amino acid</td>
<td>location</td>
<td>Getting insights into the evolution of the virus SARS-CoV-2 for genomic surveillance: hypothesis generation tool.</td>
</tr>
<tr>
<td>ViruClust</td>
<td>GISAID</td>
<td>lineage</td>
<td>date; location</td>
<td>Performing comparisons of SARS-CoV-2 genomic sequences and lineages in space and time, with the integration of different types of functional annotations; monitoring the evolution of SARS-CoV-2, facilitating the identification of variants or mutations of potential concern.</td>
</tr>
<tr>
<td>VariantHunter</td>
<td>Nextstrain/GenBank</td>
<td>lineage; amino acid</td>
<td>4-week period; location</td>
<td>Running analysis of amino acid changes frequency to observe interesting variant trends or identify novel emerging variants.</td>
</tr>
</tbody>
</table>
A. Support for small-data scenarios

Unfortunately, monitoring based on big data is coarse-grained and limited in its potential. Especially new viral pathogens, which may arise unexpectedly, will not provide an ideal big-data genomic surveillance setting. It is likely that new alarming events will be more easily caught by analyzing the arrival of small batches of uncharacterized sequences with selected characteristics. The future potential of genomic surveillance stands in small data rather than in big data. Learning from the experience built on COVID-19, new efforts should work towards methods that allow handling the data of a future viral human epidemic, supporting small data-driven analysis.

Given a virus that is starting to spread in the population, a novel framework should be put in place that aims to provide several analysis modules, whose information can be used in a complementary fashion: 1) historic analysis of patterns of mutations that regularly reappear in the evolutionary history of the virus; 2) systematic annotation of conserved areas and highlighting of mutations in such conserved areas; 3) score computation to evaluate the level of concern of observed genomes; 4) identification of sites that are under positive selection; 5) prediction of possible effects on protein functions caused by arising mutations.

New approaches could be then tested on epidemic cases known worldwide, namely small batches of open data regarding COVID-19 cases, Monkeypox (2022), Zika (2015-2016), and Ebola (2013-2016). When achieved, solutions of this kind would further advance the state of the art in understanding the various facets of genomic surveillance depending on the available small (or, rather, “not so big”) data and the domain context.

B. Continuous sharing of genome sequences

The continuous and massive depositions of sequences on behalf of worldwide laboratories are of the utmost importance for surveillance. Unfortunately, in the first part of the pandemic, there were large delays between the biologi-
cal material collection and the data submission [75]. More recently, during the second half of 2022 and the first part of 2023, data sources have recently suffered a strong slowdown in submissions. Moreover, the GISAID data source – which has been the undisputed leader of SARS-CoV-2 data sharing (even with many limitations [22]) is becoming more and more under scrutiny from scientists and funders around the world [76], [77]. As observed by one of the most virologists of the pandemic, Edward Holmes of the University of Sydney, a lesson learned from the pandemic is that data sharing is the most important thing that the community can do to support the prevention and control of pandemics.

C. Completely automated methods and services

Many methods and some services for variant and recombination identification are already offered to the community. However, all of them still need to be triggered by user queries.

Completely automatic warning systems – based on the early availability of sequences that are deposited to public databases – should be proposed, taking advantage of the lessons learned in the COVID-19 scenario.

An automatic monitoring mechanism could be used in the future as a means to provide alerts to the general scientific community on the emergence of potentially dangerous mutations in selected regions or in otherwise refined clusters of collected sequences. Such a system could put at work a daily genomic surveillance paradigm for sequences presenting mutations and patterns that are already known in the literature for their behavior or effects.

D. Integration of sequences and domain-knowledge

Currently, the provided methods and services tend not to integrate knowledge that is known a priori from domain experts or previous studies. It would be important to build a continuous integration framework that feeds data analysis pipelines with known facts. From this perspective, we can consider, in parallel, a wide range of annotations that can be gathered from public databases or predicted based on similar -known- scenarios. In this sense, a systematic approach could consider separated modules, each in charge of gathering one kind of information that – when merged with other modules’ information – can support the understanding of the complex mechanisms of the virus. Three are the expected main areas of interest: 1) functional annotations of protein regions; 2) conserved genomic regions (i.e., areas that are typically not mutated); 3) positive selection events (based on the identification of mutations that – according to predictions – can confer an advantage in the evolution of the virus).

This kind of information may be predicted by tailored algorithms, extracted from public databases, or extracted from scientific literature when no organized knowledge base is available. In this direction, we worked on CovEffect [78], a deep learning-based framework to extract SARS-CoV-2 mutations/variants effects from scientific abstracts. A large language model (GPT-2) is used to predict a series of tuples from a textual abstract in the form (variation, effect, change of effect level). For example, we can extract that the mutation V367F occurring on the Spike protein leads to a virus’ infectivity that is higher than the corresponding wild-type virus (i.e., the one not showing the mutation).

E. Modeling other epidemiological aspects

For many research questions (e.g., estimating the relative transmission rates of SARS-CoV-2 variant of concern and variants of interest [79]), observing frequencies of mutations co-occurring on genomes is not sufficient. In order to capture the rich epidemiological behavior of SARS-CoV-2, additional information is required, such as confirmed case data or the dependency of mutations (and the related raised variants) upon the demography of the geographic territories where they occurred.

VI. Future directions

A. Different geographical scales

Future-generation genomic surveillance services should tackle different geographical scales:

1) Regional settings, which need to properly monitor epidemics in local territories and organize alerting methods for bigger (country-level) organizations. An example is that of Lombardy, the region of Italy where the COVID-19 pandemic originally spread [80]. Another example is brought by the Brazilian states of the Amazonas and Ceará [81], [82].

2) Low-resources settings. The WHO has recently analyzed the context of a number of African countries where attempts to monitor COVID-19 and SARS-CoV-2 evolution have been undermined by understaffed infrastructures and a lack of resources [83].

3) Worldwide setting. This is by default monitored by the World Health Organization, with several tools available to the public, including the COVID-19 dashboard [51], periodic bulletins, and the institution of the WHO Technical Advisory Group on Virus Evolution (TAG-VE) in November 2021 [84], which has the goal of developing and implementing a global risk-monitoring framework for SARS-CoV-2 variants, based on a multidisciplinary approach.

B. Continuous reporting

Future-generation genomic surveillance services should be supported by data warehousing tools capable of producing continuous summarization, with expressive power ranging from descriptive statistics to complex data mining tasks, concerned for example with the distribution of mutations over relevant locations of the virus.

Dynamic semi-automatic learning approaches should build reactive reports on sequence characteristics and distribution, important for clinics, hospitals, or triage centers. They would provide quasi-real-time feedback to respond to relevant clinical questions generated by the virus research community: What are the infection clusters? Where are isolates with specific
characteristics coming from? Which are the most common variants? What can we say about co-occurring variants? In short, by looking at a few new sequences, it would be possible to already anticipate (or, even, predict) the implications of its spread and the variant’s representativeness in the observed population.

C. Support for other viruses and hosts

Past experience with emerging infectious diseases in the last decade, such as SARS, MERS, Zika, and Ebola, has demonstrated the importance of genomic data to study the evolution of pathogens. The most relevant work and methods have been developed for studying SARS-CoV-2, given the rapidity of its spread and the variant’s representativeness in the observed population.

In short, there exists a wealth of data and research questions on which future approaches can be tested. Genomic surveillance services should be used to monitor the spread of viruses also in hosts different from humans, as it cannot be excluded that pandemics could turn into panzootics.

REFERENCES


Bioinformatics

C. Chen, S. Nadeau, M. Yared, P. Voinov, N. Xie, C. Roemer, and

Q. Huang, Q. Zhang, P. W. Bible, Q. Liang, F. Zheng, Y. Wang, Y. Hao,

D. Mercatelli, L. Triboli, E. Fornasari, F. Ray, and F. M. Giorgi,


