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 surveillance 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 uttermost 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 typical 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-CoV2 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 [40]. A number of studies have described typical SARS-CoV-2 mutational profiles across different countries and regions [41], proposing statistical indicators for location-based mutation evolution [42] and observing changes that become recurrently prevalent in different locations, thus suggesting selective advantages [27]. 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

TABLE I
OVERVIEW OF DATA-DRIVEN METHODS FOR VIRAL VARIANT IDENTIFICATION

Proponent	Study description
Yang et al. [32] 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. [33] 2021	*Methods:* Phenetic method based on the similarity of groups according to their observable phenetic attributes. Complementary to existing methods for facilitating the identification of variants in the viral genome. *Results:* Identified 22 distinct haplogroups, gathered in four major macro haplogroups. *Highlight:* Suggested that the emergence of novel types is unlikely to be driven by convergent evolution and independent fixation of advantageous substitutions, or by a selection of recombined strains.
Chiara et al. [34] 2023	*Methods:* Automated/reproducible way to map genetic diversity in time and across different geographic regions. *Results:* Captured highly biased geographic distributions (in a complementary way w.r.t. current SARS-CoV-2 nomenclature standards).
Abe et al. [35] 2021	*Methods:* Unsupervised machine-learning method based on a batch-learning self-organizing map (BLSOM) for oligonucleotide composition. *Results:* Separation of lineages defined by GISAID (with phylogenetic methods) with high precision, further subdivided into clusters.
Bernasconi et al. [36] 2021	*Methods:* For all countries with sufficient data, it computed weekly counts of amino acid changes. It searched clusters of single changes time series; retained only those referring to increasing trends sufficiently different from previous weeks. *Results:* Timely association of clusters to variants of interest/concern whose characterization was publicly available. *Highlight:* First work to claim that the emergence of variants could be traced through purely data-driven methods, suggesting a possible predictive application for an early warning system.
De et al. [37] 2022	*Methods:* Machine learning algorithm for temporal clustering of the sequences, where distances were measured with the Levenshtein definition. *Results:* Emerging persistent variants (in agreement with known evidence) were identified, defined as "chains that remain stable over time" – whereas emerging variants with epidemiological interest were "branching events that occur over time". *Drawback:* Validation only performed on the Alpha variant data. *Highlight:* Preliminary prediction test, done on the AY.4.2 (known as 'Delta plus').
Huang et al. [38] 2022	*Methods:* Weighted network framework to model the frequency trajectories of mutations, without requiring prior subtype assignment. *Results:* The identified variants were positively assessed by using phylogenetic trees. *Highlight:* Convenient data representation showing that it is possible to rapidly and easily recognize variants overcoming prior viral subtyping.
Negi et al. [39] 2022	*Methods:* Cluster and network analysis, based on the intuition that mutations within clusters increased in frequency simultaneously. *Results:* Identification of worldwide rapidly spreading mutation and of region-specific groups of mutations.

applications were devoted to visualization purposes (e.g., the COVID-19 Viral Genome Analysis Pipeline [27], GESS [48], coronApp [49], and VirusViz [50]) with a focus on mutation-based exploration.

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 [51] – and GISAID, which integrates on its platform CoVizu [52] and CoVsurver [53]. Nextstrain [17] has a dedicated dashboard to interact with the Nextstrain phylogenetic Tree, available using GISAID data or open data (e.g., from GenBank). Other services were contributed by research centers: CoV-Spectrum [43], COVID-19 CG [44], and Outbreak.info [45]. At Politecnico di Milano, we implemented ViruClust [46] and VariantHunter [47]. Table II shows a selection of the tools with salient characteristics and differences.

IV. DATA-DRIVEN RECOMBINATION EVENTS IDENTIFICATION

Recombination represents a major contributor to the evolution of RNA viruses, occurring both in segmented and non-segmented ones. "Parents" and "child" are conventional terms used to refer to the strains that recombine and the resulting new strain; "donor" and "acceptor" refer to the parent strains, represented in a greater and lesser amount, respectively [54]. Recombination within different sublineages of the same virus requires co-circulation and co-infection of the same host; indeed, recombination events create chimeric genotypes between viral strains that infect the same cell. The clinical and epidemiological relevance of these new combinations is substantial as they have the potential to create genotypes with unique virulence and transmissibility characteristics.

Unlike other viruses that have emerged in the past two decades, coronaviruses have a high rate of recombination [55]. 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

TABLE II
OVERVIEW OF SERVICES FOR VARIANT EXPLORATION BASED ON BIG DATASETS OF VIRAL SEQUENCES

Service	Used data	Mutation search	Metadata search	Claimed purpose
CoV-Spectrum [43]	GISAID + Swiss	nucleotide; amino acid	date; location	Tracking of known variants; identification of new SARS-CoV-2 variants of concern.
COVID-19 CG [44]	GISAID	amino acid	date; location	Tracking SARS-CoV-2 single-nucleotide mutations and lineages, useful to study SARS-CoV-2 transmission, evolution, diagnostics, therapeutics, vaccines, and intervention tracking.
outbreak.info [45]	GISAID	lineage; amino acid	location	Getting insights into the evolution of the virus SARS-CoV-2 for genomic surveillance: hypothesis generation tool.
ViruClust [46]	GISAID	lineage	date; location	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.
VariantHunter [47]	Nextstrain/GenBank	lineage; amino acid	4-week period; location	Running analysis of amino acid changes frequency to observe interesting variant trends or identify novel emerging variants.

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 III 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.

At Politecnico di Milano, we produced a series of results starting from modeling efforts to understand the domain [11], [12], passing through visualization implementations [50] and providing effective methods [36] and services [46], [47] for variant tracking. In parallel, we have also studied how co-occurrence of specific mutations on lineages can suggest evolution directions [73]. Additionally, our perspective on the impact of Omicron mutations on target assays or vaccines has been commented on the Virological.org forum [74] and a preliminary attempt to spot recombination on pandemic scales has been concluded [72].

We defend that – despite the ongoing drop of interest of the international community toward SARS-CoV-2, possibly motivated by the end of three years of draining pandemic – we, as a community, should not leave aside the opportunities given by this abundance of data and promising computational problems.

Before, we saw that there exist methods that are mainly based on unsupervised machine learning algorithms that exploit a range of features that describe the trends of the epidemic. However, such methods present five main limitations:

- they need big amounts of data to produce statistically relevant evidence;
- they need data to be submitted within a short timeframe from the collection;
- they are still far from being completely automatic;

TABLE III
OVERVIEW OF DATA-DRIVEN METHODS FOR VIRAL VARIANT IDENTIFICATION

Proponent	Study description
VanInsberghe et al. [59] 2021	*Methods*: Novel lightweight approach for detecting genomes that are (only potentially) recombinant. *Results:* Estimation of the % of recombinant circulating viruses ($\sim$0.2-2.5%). Tested on database updated on February 16th, 2021, when no relevant recombinant lineage had been spotted yet.
Muller et al. [67] 2022	*Methods:* Bayesian framework (Markov chain Monte Carlo approach) to infer recombination networks (for all coronaviruses). *Results:* Showed that recombination is extremely common in the evolutionary history of SARS-like coronaviruses. First approach setting the basis for Bayesian phylogenetic tracking.
Ignatieva et al. [61] 2022	*Methods:* Parsimony-based method (KwARG) to reconstruct possible genealogical histories for samples of SARS-CoV-2 sequences. It allowed for pinpointing specific recombination events that could have generated the data. *Results:* Estimated the minimal number of recurrent mutations required to explain the data set (containing 228 sequences collected between November 2020 and February 2021) in the absence of recombination.
Preska Steinberg et al. [68] 2023	*Methods:* Adaptation of the non-phylogenetic, computationally-efficient `mcorr` method (originally developed for the analysis of bacterial genomes) to infer the parameters of homologous recombination for SARS-like coronaviruses. *Results:* Inference of recombination rates for unsampled viral reservoirs.
Zhou et al. [69] 2023	*Methods:* Information theory approach (VirusRecom) for viral recombination analysis, not relying on specific mutation sites. A recombination event was intended as a transmission process of "information" that can be accounted for by using a weighted information content to quantify the contribution of recombination to a given region on the viral genome. *Results:* Shown on simulated data and a few recombinant lineages (namely, XD, XE, and XF), with no information on how it would apply to pandemic-scale data. *Service:* Available as Python source code on GitHub.
Turkahia et al. [70] 2022	*Methods:* Approach (RIPPLES) based on breaking the potential recombinant sequence into distinct segments and replacing each of them onto a global phylogeny using maximum parsimony; the reported donor and acceptor are those that result in the highest parsimony score improvement relative to the original placement on the global phylogenetic tree. *Results:* Run on the phylogeny of May 2021, RIPPLES discovered 223 recombination events within branches of the same Pango lineages and 366 inter-lineage recombination events. These recombinants indicated approximately 2.7% of the sequenced SARS-CoV-2 genomes belonged to the detectable recombinant lineages. Tested on simulated samples and on the recombinants identified in [60]. *Service:* Implemented in the RIVET tool [71].
Alfonsi et al. [72] 2023	*Methods:* Approach based on the frequency of nucleotide mutations occurring within lineages (as defined by a reference nomenclature or otherwise determined clusters) and within all the genomes of a given sequence collection. These are used to score viral genomes by means of a likelihood-based approach and detect recombinant sequences of two lineages. *Results:* Recombinant SARS-CoV-2 genomes (or lineages) with one or two breakpoints are recognized with high accuracy, within reduced turn-around times and small discrepancies with respect to the expert manually-curated standard nomenclature.

- they disregard the possibility of interoperating sequence and mutation data with known characteristics of the virus;
- they consider frequencies of mutations co-occurring on genomes, but no other epidemiological factors.

In the following, we outline the open challenges and try to motivate the interest of the research community that designs big data services.

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 uttermost 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 limitaitons [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 data and information available for this virus. However, several data resources are also available for other kinds of viruses, as reviewed by [85], with NCBI Virus [86] and NIH Viral Genomes [87] being the most valuable and worldwide employed resources. GISAID itself has recently expanded to monkeypox with the EpiPox™ database.

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 [88], as it cannot be excluded that pandemics could turn into panzootics [89].

References

[1] S. C. Schuster, "Next-generation sequencing transforms today's biology," *Nature methods*, vol. 5, no. 1, pp. 16–18, 2008.

[2] National Human Genome Research Institute, "Dna sequencing costs: Data," 2023, last accessed: June 12th, 2023. [Online]. Available: https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

[3] J. L. Gardy and N. J. Loman, "Towards a genomics-informed, real-time, global pathogen surveillance system," *Nature Reviews Genetics*, vol. 19, pp. 9–20, 2018.

[4] World Health Organization, "Global genomic surveillance strategy for pathogens with pandemic and epidemic potential 2022–2032," 2023, last accessed: June 12th, 2023. [Online]. Available: https://www.who.int/initiatives/genomic-surveillance-strategy

[5] ——, "Terms of reference for the technical advisory group on sars-cov-2 virus evolution (tag-ve)," 2023, last accessed: June 12th, 2023. [Online]. Available: https://www.who.int/publications/m/item/terms-of-reference-for-the-technical-advisory-group-on-sars-cov-2-virus-evolution-(tag-ve)

[6] Centers for Disease Control and Prevention, "Cdc current outbreak list," 2023, last accessed: June 12th, 2023. [Online]. Available: https://www.cdc.gov/outbreaks/

[7] World Health Organization, "Tracking SARS-CoV-2 variants," 2023, last accessed: June 12th, 2023. [Online]. Available: https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/

[8] A. Maxmen, "One million coronavirus sequences: popular genome site hits mega milestone," *Nature*, vol. 593, p. 21, 2021.

[9] Y. Shu and J. McCauley, "GISAID: Global initiative on sharing all influenza data–from vision to reality," *Eurosurveillance*, vol. 22, no. 13, 2017.

[10] E. W. Sayers, M. Cavanaugh, K. Clark, K. D. Pruitt, S. T. Sherry, L. Yankie, and I. Karsch-Mizrachi, "Genbank 2023 update," *Nucleic Acids Research*, vol. 51, no. D1, pp. D141–D144, 2023.

[11] T. Alfonsi, R. Al Khalaf, S. Ceri, and A. Bernasconi, "CoV2K model, a comprehensive representation of SARS-CoV-2 knowledge and data interplay," *Scientific Data*, vol. 9, p. 260, 2022.

[12] A. Bernasconi, A. Canakoglu, P. Pinoli, and S. Ceri, "Empowering virus sequence research through conceptual modeling," in *Conceptual Modeling: 39th International Conference, ER 2020, Vienna, Austria, November 3–6, 2020, Proceedings 39*. Springer, 2020, pp. 388–402.

[13] A. S. Lauring and E. B. Hodcroft, "Genetic variants of SARS-CoV-2—what do they mean?" *Jama*, vol. 325, no. 6, pp. 529–531, 2021.

[14] Z. Yang and B. Rannala, "Molecular phylogenetics: principles and practice," *Nature reviews genetics*, vol. 13, no. 5, pp. 303–314, 2012.

[15] E. V. Koonin, V. V. Dolja, M. Krupovic, A. Varsani, Y. I. Wolf, N. Yutin, F. M. Zerbini, and J. H. Kuhn, "Global organization and proposed megataxonomy of the virus world," *Microbiology and molecular biology reviews*, vol. 84, no. 2, pp. e00 061–19, 2020.

[16] A. Rambaut, E. C. Holmes, A. O'Toole, V. Hill, J. T. McCrone, C. Ruis, L. du Plessis, and O. G. Pybus, "A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology," *Nature Microbiology*, vol. 5, no. 11, pp. 1403–1407, 11 2020.

[17] J. Hadfield, C. Megill, S. M. Bell, J. Huddleston, B. Potter, C. Callender, P. Sagulenko, T. Bedford, and R. A. Neher, "Nextstrain: real-time tracking of pathogen evolution," *Bioinformatics*, vol. 34, no. 23, pp. 4121–4123, 12 2018.

[18] E. Callaway, "Coronavirus variants get Greek names – but will scientists use them?" *Nature*, vol. 594, p. 162, 2021.

[19] G. T. Ryan, "Tweet on twitter account," 2023, last accessed: June 12th, 2023. [Online]. Available: https://twitter.com/TRyanGregory/status/1574055652649041921?s=20

[20] C. Yue, W. Song, L. Wang, F. Jian, X. Chen, F. Gao, Z. Shen, Y. Wang, X. Wang, and Y. Cao, "Ace2 binding and antibody evasion in enhanced transmissibility of xbb.1.5," *The Lancet Infectious Diseases*, vol. 23, no. 3, pp. 278–280, 2023.

[21] D. Yamasoba, K. Uriu, A. Plianchaisuk, Y. Kosugi, L. Pan, J. Zahradnik, J. Ito, and K. Sato, "Virological characteristics of the sars-cov-2 omicron xbb.1.16 variant," *The Lancet Infectious Diseases*, vol. 23, no. 6, pp. 655–656, 2023.

[22] A. Bernasconi, A. Canakoglu, M. Masseroli, P. Pinoli, and S. Ceri, "A review on viral data sources and search systems for perspective mitigation of covid-19," *Briefings in bioinformatics*, vol. 22, no. 2, pp. 664–675, 2021.

[23] International Nucleotide Sequence Database Collaboration, "Homepage," 2023, last accessed: June 12th, 2023. [Online]. Available: http://www.insdc.org

[24] N. A. O'Leary, M. W. Wright, J. R. Brister, S. Ciufo, D. Haddad, R. McVeigh, B. Rajput, B. Robbertse, B. Smith-White, D. Ako-Adjei *et al.*, "Reference sequence (refseq) database at ncbi: current status, taxonomic expansion, and functional annotation," *Nucleic acids research*, vol. 44, no. D1, pp. D733–D745, 2016.

[25] C. Cummins, A. Ahamed, R. Aslam, J. Burgin, R. Devraj, O. Edbali, D. Gupta, P. W. Harrison, M. Haseeb, S. Holt *et al.*, "The European Nucleotide Archive in 2021," *Nucleic acids research*, vol. 50, no. D1, pp. D106–D110, 2022.

[26] The COVID-19 Genomics UK Consortium, "An integrated national scale SARS-CoV-2 genomic surveillance network," *The Lancet Microbe*, vol. 1, no. 3, pp. E99–E100, 2020.

[27] B. Korber, W. M. Fischer, S. Gnanakaran, H. Yoon, J. Theiler, W. Abfalterer, N. Hengartner, E. E. Giorgi, T. Bhattacharya, B. Foley, K. M. Hastie, M. D. Parker, D. G. Partridge, C. M. Evans, T. M. Freeman, T. I. de Silva, A. Angyal, R. L. Brown, L. Carrilero, L. R. Green, D. C. Groves, K. J. Johnson, A. J. Keeley, B. B. Lindsey, P. J. Parsons, M. Raza, S. Rowland-Jones, N. Smith, R. M. Tucker, D. Wang, M. D. Wyles, C. McDanal, L. G. Perez, H. Tang, A. Moon-Walker, S. P. Whelan, C. C. LaBranche, E. O. Saphire, and D. C. Montefiori, "Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus," *Cell*, vol. 182, no. 4, pp. 812–827, 2020.

[28] E. B. Hodcroft, M. Zuber, S. Nadeau, T. G. Vaughan, K. H. D. Crawford, C. L. Althaus, M. L. Reichmuth, J. E. Bowen, A. C. Walls, D. Corti, J. D. Bloom, D. Veesler, D. Mateo, A. Hernando, I. Comas, F. G. Candelas, S.-S. Consortium, T. Stadler, and R. A. Neher, "Spread of a SARS-CoV-2 variant through Europe in the summer of 2020," *Nature*, vol. 595, p. 707–712, 2021.

[29] K. Ziegler, P. Steininger, R. Ziegler, J. Steinmann, K. Korn, and A. Ensser, "SARS-CoV-2 samples may escape detection because of a single point mutation in the N gene," *Eurosurveillance*, vol. 25, no. 39, p. 2001650, 2020.

[30] R. Wang, Y. Hozumi, C. Yin, and G.-W. Wei, "Mutations on COVID-19 diagnostic targets," *Genomics*, vol. 112, no. 6, pp. 5204–5213, 2020.

[31] S. A. Madhi, V. Baillie, C. L. Cutland, M. Voysey, A. L. Koen, L. Fairlie, S. D. Padayachee, K. Dheda, S. L. Barnabas, Q. E. Bhorat *et al.*, "Efficacy of the ChAdOx1 nCoV-19 Covid-19 vaccine against the

B.1.351 variant," *New England Journal of Medicine*, vol. 384, no. 20, pp. 1885–1898, 2021.

[32] H.-C. Yang, C.-h. Chen, J.-H. Wang, H.-C. Liao, C.-T. Yang, C.-W. Chen, Y.-C. Lin, C.-H. Kao, M.-Y. J. Lu, and J. C. Liao, "Analysis of genomic distributions of sars-cov-2 reveals a dominant strain type with strong allelic associations," *Proceedings of the National Academy of Sciences*, vol. 117, no. 48, pp. 30 679–30 686, 2020.

[33] M. Chiara, D. S. Horner, C. Gissi, and G. Pesole, "Comparative Genomics Reveals Early Emergence and Biased Spatiotemporal Distribution of SARS-CoV-2," *Molecular Biology and Evolution*, vol. 38, no. 6, pp. 2547–2565, 2021a.

[34] M. Chiara, D. S. Horner, E. Ferrandi, C. Gissi, and G. Pesole, "Haplocov: unsupervised classification and rapid detection of novel emerging variants of sars-cov-2," *Communications Biology*, vol. 6, no. 1, p. 443, 2023.

[35] T. Abe, R. Furukawa, Y. Iwasaki, and T. Ikemura, "Time-series trend of pandemic sars-cov-2 variants visualized using batch-learning self-organizing map for oligonucleotide compositions," *Data Science Journal*, vol. 20, no. 1, 2021.

[36] A. Bernasconi, L. Mari, R. Casagrandi, and S. Ceri, "Data-driven analysis of amino acid change dynamics timely reveals SARS-CoV-2 variant emergence," *Scientific Reports*, vol. 11, p. 21068, 2021.

[37] A. de Hoffer, S. Vatani, C. Cot, G. Cacciapaglia, M. L. Chiusano, A. Cimarelli, F. Conventi, A. Giannini, S. Hohenegger, and F. Sannino, "Variant-driven early warning via unsupervised machine learning analysis of spike protein mutations for covid-19," *Scientific Reports*, vol. 12, p. 9275, 2022.

[38] Q. Huang, Q. Zhang, P. W. Bible, Q. Liang, F. Zheng, Y. Wang, Y. Hao, and Y. Liu, "A new way to trace sars-cov-2 variants through weighted network analysis of frequency trajectories of mutations," *Frontiers in Microbiology*, vol. 13, 2022.

[39] S. S. Negi, C. H. Schein, and W. Braun, "Regional and temporal coordinated mutation patterns in sars-cov-2 spike protein revealed by a clustering and network analysis," *Scientific Reports*, vol. 12, p. 1128, 2022.

[40] J. A. Plante, B. M. Mitchell, K. S. Plante, K. Debbink, S. C. Weaver, and V. D. Menachery, "The variant gambit: COVID-19's next move," *Cell Host & Microbe*, vol. 29, no. 4, pp. 508–515, 2021.

[41] D. Mercatelli and F. M. Giorgi, "Geographic and genomic distribution of SARS-CoV-2 mutations," *Frontiers in Microbiology*, vol. 11, p. 1800, 2020.

[42] P. Troyano-Hernáez, R. Reinosa, and Á. Holguín, "Evolution of SARS-CoV-2 envelope, membrane, nucleocapsid, and spike structural proteins from the beginning of the pandemic to September 2020: a global and regional approach by epidemiological week," *Viruses*, vol. 13, no. 2, p. 243, 2021.

[43] C. Chen, S. Nadeau, M. Yared, P. Voinov, N. Xie, C. Roemer, and T. Stadler, "CoV-Spectrum: analysis of globally shared SARS-CoV-2 data to identify and characterize new variants," *Bioinformatics*, vol. 38, no. 6, pp. 1735–1737, 2022.

[44] A. T. Chen, K. Altschuler, S. H. Zhan, Y. A. Chan, and B. E. Deverman, "COVID-19 CG enables SARS-CoV-2 mutation and lineage tracking by locations and dates of interest," *Elife*, vol. 10, p. e63409, 2021a.

[45] K. Gangavarapu, A. A. Latiff, J. L. Mullen, M. Alkuzweny, E. Hufbauer, G. Tsueng, E. Haag, M. Zeller, C. M. Aceves, K. Zaiets *et al.*, "Outbreak.info genomic reports: scalable and dynamic surveillance of SARS-CoV-2 variants and mutations," *medRxiv*, 2022. [Online]. Available: https://doi.org/10.1101/2022.01.27.22269965

[46] L. Cilibrasi, P. Pinoli, A. Bernasconi, A. Canakoglu, M. Chiara, and S. Ceri, "Viruclust: direct comparison of sars-cov-2 genomes and genetic variants in space and time," *Bioinformatics*, vol. 38, no. 7, pp. 1988–1994, 2022.

[47] P. Pinoli, Canakoglu, S. Ceri, M. Chiara, E. Ferrandi, L. Minotti, and A. Bernasconi, "Genosurf: metadata driven semantic search system for integrated genomic datasets," *Database*, vol. 2023 *[Accepted for publication]*, 2023.

[48] S. Fang, K. Li, J. Shen, S. Liu, J. Liu, L. Yang, C.-D. Hu, and J. Wan, "GESS: a database of global evaluation of SARS-CoV-2/hCoV-19 sequences," *Nucleic Acids Research*, vol. 49, no. D1, pp. D706–D714, 2020.

[49] D. Mercatelli, L. Triboli, E. Fornasari, F. Ray, and F. M. Giorgi, "Coronapp: A web application to annotate and monitor SARS-CoV-2 mutations," *Journal of Medical Virology*, vol. 93, no. 5, pp. 3238–3245, 2020.

[50] A. Bernasconi, A. Gulino, T. Alfonsi, A. Canakoglu, P. Pinoli, A. Sandionigi, and S. Ceri, "VirusViz: comparative analysis and effective visualization of viral nucleotide and amino acid variants," *Nucleic Acids Research*, vol. 49, no. 15, p. e90, 2021a.

[51] World Health Organization, "WHO COVID-19 Dashboard," 2020, last accessed: June 12th, 2023. [Online]. Available: https://covid19.who.int/

[52] R.-C. Ferreira, E. Wong, G. Gugan, K. Wade, M. Liu, L. M. Baena, C. Chato, B. Lu, A. S. Olabode, and A. F. Poon, "Covizu: Rapid analysis and visualization of the global diversity of sars-cov-2 genomes," *Virus Evolution*, vol. 7, no. 2, p. veab092, 2021.

[53] S. Khare, C. Gurry, L. Freitas, M. B. Schultz, G. Bach, A. Diallo, N. Akite, J. Ho, R. T. Lee, W. Yeo *et al.*, "Gisaid's role in pandemic response," *China CDC weekly*, vol. 3, no. 49, p. 1049, 2021.

[54] D. Focosi and F. Maggi, "Recombination in coronaviruses, with a focus on sars-cov-2," *Viruses*, vol. 14, no. 6, p. 1239, 2022.

[55] D. Forni, R. Cagliani, M. Clerici, and M. Sironi, "Molecular evolution of human coronavirus genomes," *Trends in microbiology*, vol. 25, no. 1, pp. 35–48, 2017.

[56] M. H. Schierup and J. Hein, "Consequences of recombination on traditional phylogenetic analysis," *Genetics*, vol. 156, no. 2, pp. 879–891, 2000.

[57] D. Posada, K. A. Crandall, and E. C. Holmes, "Recombination in evolutionary genomics," *Annual Review of Genetics*, vol. 36, no. 1, pp. 75–97, 2002.

[58] R. Shiraz and S. Tripathi, "Enhanced recombination among omicron subvariants of sars-cov-2 contributes to viral immune escape," *Journal of Medical Virology*, vol. 95, no. 2, p. e28519, 2023.

[59] D. VanInsberghe, A. S. Neish, A. C. Lowen, and K. Koelle, "Recombinant sars-cov-2 genomes circulated at low levels over the first year of the pandemic," *Virus Evolution*, vol. 7, no. 2, p. veab059, 2021.

[60] B. Jackson, M. F. Boni, M. J. Bull, A. Colleran, R. M. Colquhoun, A. C. Darby, S. Haldenby, V. Hill, A. Lucaci, J. T. McCrone *et al.*, "Generation and transmission of interlineage recombinants in the sars-cov-2 pandemic," *Cell*, vol. 184, no. 20, pp. 5179–5188, 2021.

[61] A. Ignatieva, J. Hein, and P. A. Jenkins, "Ongoing recombination in sars-cov-2 revealed through genealogical reconstruction," *Molecular Biology and Evolution*, vol. 39, no. 2, p. msac028, 2022.

[62] T. Sekizuka, K. Itokawa, M. Saito, M. Shimatani, S. Matsuyama, H. Hasegawa, T. Saito, and M. Kuroda, "Genome recombination between the delta and alpha variants of severe acute respiratory syndrome coronavirus 2 (sars-cov-2)," *Japanese journal of infectious diseases*, vol. 75, no. 4, pp. 415–418, 2022.

[63] R. Duerr, H. Zhou, T. Tada, D. Dimartino, C. Marier, P. Zappile, G. Wang, J. Plitnick, S. B. Griesemer, R. Girardin, J. Machowski, S. Bialosuknia, E. Lasek-Nesselquist, S. L. Hong, G. Baele, M. Dittmann, M. B. Ortigoza, P. J. Prasad, K. McDonough, N. R. Landau, K. St George, and A. Heguy, "Delta-omicron recombinant escapes therapeutic antibody neutralization," *iScience*, vol. 26, no. 2, p. 106075, 2023.

[64] F. Kreier *et al.*, "Deltacron: the story of the variant that wasn't," *Nature*, vol. 602, no. 7895, p. 19, 2022.

[65] H. M. Lam, O. Ratmann, and M. F. Boni, "Improved algorithmic complexity for the 3seq recombination detection algorithm," *Molecular biology and evolution*, vol. 35, no. 1, pp. 247–251, 2018.

[66] D. P. Martin, A. Varsani, P. Roumagnac, G. Botha, S. Maslamoney, T. Schwab, Z. Kelz, V. Kumar, and B. Murrell, "Rdp5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets," *Virus Evolution*, vol. 7, no. 1, p. veaa087, 2021.

[67] N. F. Müller, K. E. Kistler, and T. Bedford, "A bayesian approach to infer recombination patterns in coronaviruses," *Nature communications*, vol. 13, no. 1, p. 4186, 2022.

[68] A. Preska Steinberg, O. K. Silander, and E. Kussell, "Correlated substitutions reveal sars-like coronaviruses recombine frequently with a diverse set of structured gene pools," *Proceedings of the National Academy of Sciences*, vol. 120, no. 5, p. e2206945119, 2023.

[69] Z.-J. Zhou, C.-H. Yang, S.-B. Ye, X.-W. Yu, Y. Qiu, and X.-Y. Ge, "Virusrecom: an information-theory-based method for recombination detection of viral lineages and its application on sars-cov-2," *Briefings in Bioinformatics*, vol. 24, no. 1, p. bbac513, 2023.

[70] Y. Turakhia, B. Thornlow, A. Hinrichs, J. McBroome, N. Ayala, C. Ye, K. Smith, N. De Maio, D. Haussler, R. Lanfear *et al.*, "Pandemic-scale phylogenomics reveals the sars-cov-2 recombination landscape," *Nature*, vol. 609, no. 7929, pp. 994–997, 2022.

[71] K. Smith, C. Ye, and Y. Turakhia, "Tracking and curating putative sars-cov-2 recombinants with rivet," *bioRxiv*, pp. 2023–02, 2023.

[72] T. Alfonsi, A. Bernasconi, M. Chiara, and S. Ceri, "Data-driven recombination detection in viral genomes," *bioRxiv*, 2023. [Online]. Available: https://doi.org/10.1101/2023.06.05.543733

[73] R. Al Khalaf, A. Bernasconi, P. Pinoli, and S. Ceri, "Analysis of co-occurring and mutually exclusive amino acid changes and detection of convergent and divergent evolution events in SARS-CoV-2," *Computational and Structural Biotechnology Journal*, vol. 20, pp. 4238–4250, 2022.

[74] A. Bernasconi, P. Pinoli, R. Al Khalaf, T. Alfonsi, A. Canakoglu, L. Cilibrasi, and S. Ceri, "Report on omicron spike mutations on epitopes and immunological/epidemiological/kinetics effects from literature," 2021, last accessed: June 12th, 2023. [Online]. Available: https://virological.org/t/report-on-omicron-spike-mutations-on-epitopes-and-immunological-epidemiological-kinetics-effects-from-literature/770

[75] K. Kalia, G. Saberwal, and G. Sharma, "The lag in SARS-CoV-2 genome submissions to GISAID," *Nature Biotechnology*, vol. 39, no. 9, pp. 1058–1060, 2021.

[76] M. Lenharo, "Gisaid in crisis: can the controversial covid genome database survive?" *Nature*, vol. 617, no. 7961, pp. 455–457, 2023.

[77] S. Mallapaty, "Covid-origins report sparks debate over major genome hub gisaid." *Nature*, pp. 13–14, 2023.

[78] G. Serna García, R. Al Khalaf, F. Invernici, S. Ceri, and A. Bernasconi, "Coveffect: interactive system for mining the effects of sars-cov-2 mutations and variants based on deep learning," *GigaScience*, vol. 12, p. giad036, 2023.

[79] M. D. Figgins and T. Bedford, "Sars-cov-2 variant dynamics across us states show consistent differences in effective reproduction numbers," *medRxiv*, 2021. [Online]. Available: https://doi.org/10.1101/2021.12.09.21267544

[80] D. Cereda, M. Manica, M. Tirani, F. Rovida, V. Demicheli, M. Ajelli, P. Poletti, F. Trentini, G. Guzzetta, V. Marziano *et al.*, "The early phase of the COVID-19 epidemic in Lombardy, Italy," *Epidemics*, vol. 37, p. 100528, 2021.

[81] F. G. Naveca, V. Nascimento, V. C. de Souza, A. de Lima Corado, F. Nascimento, G. Silva, Á. Costa, D. Duarte, K. Pessoa, M. Mejía *et al.*, "COVID-19 in Amazonas, Brazil, was driven by the persistence of endemic lineages and P.1 emergence," *Nature Medicine*, vol. 27, p. 1230–1238, 2021.

[82] F. A. da Silva Oliveira, M. V. de Holanda, L. B. Lima, M. B. Dantas, I. O. Duarte, L. G. Z. de Castro, L. L. B. de Oliveira, C. R. K. Paier, C. d. F. A. Moreira-Nunes, N. C. B. Lima *et al.*, "Genomic surveillance: Circulating lineages and genomic variation of SARS-CoV-2 in early pandemic in Ceará state, Northeast Brazil," *Virus Research*, vol. 321, p. 198908, 2022.

[83] World Health Organization, "Reflecting on the implementation of genomic surveillance for COVID-19 and beyond in the African Region," 2022, last accessed: June 12th, 2023. [Online]. Available: https://www.who.int/news/item/16-09-2022-reflecting-on-the-implementation-of-genomic-surveillance-for-COVID-19-and-beyond-in-the-african-region

[84] L. Subissi, A. von Gottberg, L. Thukral, N. Worp, B. B. Oude Munnink, S. Rathore, L. J. Abu-Raddad, X. Aguilera, E. Alm, B. N. Archer *et al.*, "An early warning system for emerging sars-cov-2 variants," *Nature medicine*, vol. 28, no. 6, pp. 1110–1115, 2022.

[85] D. Sharma, P. Priyadarshini, and S. Vrati, "Unraveling the web of viroinformatics: computational tools and databases in virus research," *Journal of Virology*, vol. 89, no. 3, pp. 1489–1501, 2015.

[86] E. L. Hatcher, S. A. Zhdanov, Y. Bao, O. Blinkova, E. P. Nawrocki, Y. Ostapchuck, A. A. Schäffer, and J. R. Brister, "Virus Variation Resource–improved response to emergent viral outbreaks," *Nucleic Acids Research*, vol. 45, no. D1, pp. D482–D490, 2017.

[87] J. R. Brister, D. Ako-Adjei, Y. Bao, and O. Blinkova, "Ncbi viral genomes resource," *Nucleic Acids Research*, vol. 43, no. D1, pp. D571–D577, 2015.

[88] N. Decaro and A. Lorusso, "Novel human coronavirus (SARS-CoV-2): A lesson from animal coronaviruses," *Veterinary Microbiology*, vol. 44, p. 108693, 2020.

[89] R. Gollakner and I. Capua, "Is COVID-19 the first pandemic that evolves into a panzootic?" *Veterinaria Italiana*, vol. 56, no. 1, pp. 11–12, 2020.