POLITECNICO MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

Variant Hunter: a tool for fast detection of emerging SARS-CoV-2 variants

Tesi di Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: **Luca Minotti**

Student ID: 963045
Advisor: Prof. Stefano Ceri
Co-advisors: Dr. Anna Bernasconi, Dr. Pietro Pinoli
Academic Year: 2021-22

Abstract

Since the first reports of coronavirus cases in China and the publication of the first SARS-CoV-2 sequence in 2019, the virus has undergone numerous mutations. Indeed, viruses change when they replicate and spread in a population, and this may give them the ability to infect more efficiently or be recognized less easily by the immune system. Monitoring changes in the genetic code of SARS-CoV-2 allows to model the present status, and predict the near-future evolution of the pandemic. This enables improvements in vaccines, testing, and adapting social distancing measures accordingly.

While more than tens of millions of genomic sequences of SARS-CoV-2 are available, their analysis would generally require a significant amount of manual work and engage a huge number of virologists worldwide. Variant Hunter moves toward automating this work. Specifically, the tool analyzes the frequencies of amino acid mutations detected in the sequences over 4-week spans in specific locations, in order to hunt novel emerging variants as early as possible.

Variant Hunter mainly supports two types of analysis, namely lineage-independent and lineage-specific analysis. The former has the primary objective of identifying the occurrence of new mutations at regional, national, or continental level. In contrast, the second feature focuses on the sequences of a specific lineage to support the discovery of new sub-lineages. Variant Hunter is based on simple yet effective statistics and visual representations. A Diffusion Heatmap allows to observe the trends in a quick way. The tool also provides Diffusion Trend charts and Odd Ratio plots. The former depict the evolution of the mutation frequencies over time, while the latter provide a direct comparison of the spread over the weeks.

Variant Hunter is implemented in the form of a web application that runs in standard browsers. The large amount of data it has to handle have required considerable design effort. Much of the information and intermediate results are pre-computed so as to make a typical analysis running in less than a second on a common computer.

Thanks the collaboration of several virologists in different parts of the world, the tool was designed to meet the specific needs of the field. Thus, the simple and intuitive design of

Variant Hunter is combined with the high flexibility brought about by the large number of options available.

In addition, research institutions also have the possibility to use the tool to analyze restricted-access or other private sequencing data by installing the Docker version. In particular, the latter is rich of configuration options that allow to save time and computational resources.

The goodness of the work is also confirmed by the fact that the appearance of a large number of variants could have been easily predicted through the tool's features.

Keywords: SARS-Cov-2, COVID-19, lineages, genomic surveillance, variant tracking, variant discovery

Abstract in lingua italiana

Dalle prime segnalazioni di casi di coronavirus in Cina e dalla pubblicazione della prima sequenza di SARS-CoV-2 nel 2019, il virus ha subito numerose mutazioni. In generale, i virus cambiano quando si replicano e si diffondono in una popolazione, e questo può dare loro la capacità di infettare in modo più efficiente o di essere riconosciuti meno facilmente dal sistema immunitario. Il monitoraggio dei cambiamenti nel codice genetico del SARS-CoV-2 permette di modellare lo stato attuale e di prevedere l'evoluzione della pandemia. Ciò consente di migliorare i vaccini, i test e di adattare di conseguenza le misure di distanziamento sociale.

Sebbene siano disponibili più di decine di milioni di sequenze genomiche del SARS-CoV-2, la loro analisi richiede generalmente una notevole quantità di lavoro manuale e coinvolge un numero considerevole di virologi in tutto il mondo. Variant Hunter mira ad automatizzare questo lavoro. In particolare, lo strumento analizza le frequenze delle mutazioni aminoacide rilevate nelle sequenze nell'arco di 4 settimane in luoghi specifici, al fine di individuare le nuove varianti emergenti il più presto possibile.

Variant Hunter supporta principalmente due tipi di analisi: quella lignaggio-indipendente e quella lignaggio-specifica. La prima ha l'obiettivo primario di identificare la presenza di nuove mutazioni a livello regionale, nazionale o continentale. La seconda, invece, si concentra sulle sequenze di uno specifico lignaggio per favorire la scoperta di nuovi sotto-lignaggi. Variant Hunter si basa su statistiche e rappresentazioni visive semplici ma efficaci. Una heatmap di diffusione permette di osservare le tendenze in modo rapido. Lo strumento fornisce anche grafici del trend di diffusione e di odd ratio. I primi rappresentano l'evoluzione delle frequenze delle mutazioni nel tempo, mentre i secondi forniscono un confronto diretto della diffusione nelle settimane.

Variant Hunter è implementato sotto forma di applicazione web. La grande quantità di dati che il sistema deve gestire ha richiesto un notevole sforzo di progettazione. Gran parte delle informazioni e dei risultati intermedi sono precalcolati, in modo da rendere un'analisi tipica eseguibile in meno di un secondo su un comune computer.

Grazie alla collaborazione di alcuni virologi in diverse parti del mondo, lo strumento è stato è stato disegnato per soddisfare le esigenze specifiche del settore. Il design semplice e intuitivo di Variant Hunter si combina quindi con l'elevata flessibilità garantita dal gran numero di opzioni disponibili.

Inoltre, gli istituti di ricerca hanno anche la possibilità di utilizzare lo strumento per analizzare dati di sequenziamento ad accesso limitato o privati, installando la versione Docker. In particolare, quest'ultima è ricca di opzioni di configurazione che permettono di risparmiare tempo e risorse computazionali.

La bontà del lavoro è confermata anche dal fatto che la comparsa di un gran numero di varianti poteva essere facilmente prevista grazie alle caratteristiche dello strumento.

Parole chiave: SARS-Cov-2, COVID-19, lignaggio, sorveglianza genomica, tracciamento varianti, scoperta varianti

Contents

1 | Introduction

1.1. Scenario and Problem Definition

Since its emergence in December 2019, *Severe Acute Respiratory Syndrome Coronavirus 2* (SARS-CoV-2) has had an enormous impact on the healthcare system worldwide. New estimates from the *World Health Organization* (WHO) show that the total number of deaths associated directly or indirectly with the COVID-19 pandemic (described as "*excess mortality*") between 1 January 2020 and 31 December 2021 was approximately 14.9 million [30].

Careful monitoring of emerging variants is of paramount importance in the fight against COVID-19. Indeed, all viruses change as they reproduce and spread in a population. Every time SARS-CoV-2 replicates, there is an opportunity for the virus to change (*mutate*). Most of the mutations do not alter the virus's properties because they do not affect the major proteins involved in infection and transmission. However, some changes do impact virus's characteristics, such as how easily it spreads, the associated disease severity, or the performance of vaccines, therapeutic medicines, diagnostic and sequencing tools, or other public health and social measures [31]. Through *genomic surveillance*, scientists and virologists track the spread of variants and monitor changes to the genetic code of SARS-CoV-2.

Genomics surveillance represents the first line of defense against the spreading of novel viral variants, as it allows to obtain statistics that model the present status and predict the near-future evolution of the pandemic. Vaccination programs and testing can be improved on the basis of regularly updated reports of the emerging variants, including updating future vaccines if required. Detection of new variants associated with increased infectivity or more severe disease, support prevention efforts and strengthens the health institutions response [3].

1.2. Aim and Idea

This work is positioned in the context described above and represents a first step toward simplifying mutation analysis of SARS-CoV-2. More specifically, the idea behind Variant Hunter is to analyze the frequencies of SARS-CoV-2 amino acid mutations in order to observe interesting variant trends or identify new emerging variants.

The main goal of this project is to provide an highly flexible and user-friendly tool to support monitoring of viral evolution, empowering genomic surveillance.

The system, developed as a web app, provides visualizations and uses statistics carefully designed to highlight patterns and make them identifiable at a glance.

Variant Hunter is structured into two main analyses.

- The first is a lineage-independent analysis and focuses on the examination of all amino acid substitutions recorded in a given period and location.
  The primary objective here is to identify the occurrence of new mutations at regional, national, or continental level.

- In contrast, the second feature focuses on the sequences of a specific lineage.
  The aim in this case is to simplify the discovery of new sub-lineages of a parent one, arising from the occurrence of mutations that are atypical for the lineage under consideration. Data can be analyzed at different levels of granularity (i.e., regional, national or continental) and for specific time periods and locations.

Ideally, a typical workflow consists, at first, on the analysis of the data in a lineage-independent manner. From the latter, the user should be able to identify mutations that are clearly increasing in prevalence and also to extract with which lineage the amino acid substitutions of interest are mostly associated. This will allow for more in-depth examinations using linear-specific analyses.

1.3. Outline

This document is structured as follows:

- ***Chapter 2:*** *Virology Primer* – A general introduction to virology presenting basic concepts of genetics and viruses.

- ***Chapter 3:*** *Tool Description* – A detailed description of the software that presents its features providing, for each of them, meaningful use cases. In particular, this chapter focuses on the two possible analyses: lineage-specific and lineage-independent.

An extension of the tool that facilitates the exploration of the data set will also be discussed.

- ***Chapter 4:*** *Design and Implementation* – A precise discussion of the software design and major implementation aspects. This section details the steps through which the solution has been developed: starting from the analysis of the available sequence metadata to the introduction of the implemented analysis techniques.
- ***Chapter 5:*** *Deployment* – This chapter presents how the software was deployed. In particular, it discusses the possibility of running the tool through Docker, enabling its usage also with private sequencing data.
- ***Chapter 6:*** *Performances and Testing* – Discussion of the performances the system is capable of and the steps taken to improve them.
- ***Chapter 7:*** *Conclusions and Future Developments* – This section draws conclusions from the experiment and briefly presents some possible extensions of the tool.

2 | Virology Primer

The purpose of this chapter is to introduce the research area covered by this document by briefly presenting its main concepts. In particular, some basic concepts of genetics and viruses that are particularly relevant to the analysis carried out are discussed in the following sections.

2.1. Introduction to Viruses

Viruses are simple microorganisms consisting of: a nucleic acid (DNA or RNA depending on the type of virus), which contains the genetic information necessary for their multiplication; and a variable number of proteins, some of them located on the outer envelope (called the viral capsid). Viruses are not autonomous and they are only able to live and multiply within the cells of their host organism (e.g., humans) [15],[23].

Coronaviruses (CoVs), so-called because of the crown-shaped spikes on their surface, are a large family of RNA (ribonucleic acid) respiratory viruses that can cause mild to severe illness, from the common cold to respiratory syndromes such as MERS (Middle East respiratory syndrome) and SARS (Severe acute respiratory syndrome) [14].

Ribonucleic acid (abbreviated as RNA) is a nucleic acid found in all living cells that exhibits structural similarities to DNA. However, in contrast to DNA, RNA is often single-stranded [17].

In general, an RNA molecule has a backbone consisting of alternating phosphate groups and the sugar ribose. Attached to each sugar is one of four bases: adenine (A), uracil (U), cytosine (C) or guanine (G) [17].

Coronaviruses cause infections in humans and various animals, including birds and mammals such as camels, cats, and bats. They are able to infect different species, and this "species jumping" occurs through changes (mutations) in the virus' genetic heritage that make it capable of infecting (adapting to) new animal species, including humans.

The new 2019 coronavirus (officially identified as SARS-CoV-2) is a new strain, which appeared in late 2019 in Wuhan, China that causes a severe acute respiratory syndrome called COVID-19 (COronaVIrus Disease-2019) by the World Health Organization [14].

2.2. Virus transmission and Variants

Upon infection, the virus enters the cell using one or more proteins on its outer envelope (in the case of SARS-CoV-2 mainly the Spike protein). Once inside, it releases its genetic material (called genome), containing all the information needed to replicate itself, in the infected cell and cause the latter to produce many new viruses, copies of itself. The latter will not only infect other cells of the same individual, but will also be transmitted to other people [15],[23].

The most critical point is that all viruses change when they replicate and spread in a population. In general, RNA viruses, such as SARS-CoV-2 and influenza, mutate much faster than DNA viruses [15]. Each time SARS-CoV-2 replicates, the virus has a chance to change.

More precisely, we define a **mutation** as a single change in the virus genome [5].
In contrast, a **variant** is defined as the set of one or more mutations that differentiate it from other "versions" of the SARS-CoV-2 virus [4].
Finally, a group of closely genetically related viral variants, derived from a common ancestor, is called a **lineage** [5]. The Pango nomenclature [19] is being used by researchers and public health agencies worldwide to distinguish the various lineages.

Many mutations do not affect the virus' ability to spread or cause disease, because they do not alter the major proteins involved in infection and transmission. However, when one of these mutations affects the virus' ability to spread or cause disease, a competitive advantage over other SARS-CoV-2 lineages may occur [31].
This is especially the case when the changes affect proteins on the outer envelope (in the case of SARS-CoV-2 the Spike protein) of the virus. Indeed, these proteins are the ones that the host immune system sees first and against which it triggers a stronger immune response, including antibody production.
Consequently, changing the characteristics of these proteins may give the virus the ability to infect more efficiently or be recognized less easily by the immune system [15]. In those cases, close monitoring of the new, potentially dangerous variant is essential.

When a lineage or group of lineages exhibits characteristics that have a non-negligible impact on public health, they are classified as "variant of interest" or "variant of concern" by the WHO (World Health Organization) [31]. More specifically:

Variants of interest (VOI) : a SARS-CoV-2 variant demonstrated to be associated with both:

- genetic changes that are expected or known to affect virus characteristics, such as transmissibility, disease severity, immune evasion, diagnostic or therapeutic evasion;
- significant community transmission or multiple clusters of COVID-19, in multiple countries, with increasing relative prevalence and increasing numbers of cases over time, or other obvious epidemiological impacts that suggest an emerging global public health risk [31].

Variants of concern (VOC) : a SARS-CoV-2 variant associated with one or more of the following changes at a degree of global public health significance:

- increased transmissibility or remarkable change in the epidemiology of COVID-19;
- increased virulence or change in clinical presentation of the disease;
- decreased effectiveness of social and public health measures or available diagnostic, vaccine and therapeutic products [31].

Variants under monitoring (VUM) : a SARS-CoV-2 variant associated with genetic modifications suspected of affecting virus characteristics, with some indications that it may pose a future risk. However, the evidence for phenotypic or epidemiologic impact is currently unclear: enhanced monitoring and repeated evaluation is needed [31].

A variant of interest (VOI) or variant of concern (VOC) may be downgraded in this list after a significant and sustained reduction in its national and regional proportions over time, or after other evidence indicates that a variant does not pose a significant public health risk.

It should be noted, however, that such variants continue to be of interest to the scientific community and could be reclassified as VOCs or VOIs if the epidemiological situation changes in a worrisome way.

2.3. Genomic Sequencing and Surveillance

Biologists and virologists use genomic sequencing to identify the variant of SARS-CoV-2 present in a sample. In general, sequencing consists of the process of deciphering the genetic material present in an organism or virus [4].

Scientists constantly accumulate sequences and analyze the similarities and differences between these sequences in a process called genomic surveillance. Through genomic surveillance, scientists track the spread of variants by monitoring changes in the genetic code of SARS-CoV-2 [4].

Note that for this process to be effective, sequencing a sample of every COVID-19 case is not required. Instead, the goal is to collect sufficient sequence data from representative populations, to identify new variants and monitor trends in circulating mutations.

Overall, this information is used to better understand the impact of variants on public health, support prevention efforts and strengthens the healthcare institutions response. Indeed, knowing the mutations of a virus can help predict what its behavior will be: whether it will be more or less infectious, more or less resistant to vaccines or therapies, etc. In addition, knowing what variants are circulating and, more importantly, finding out if new variants are emerging is of paramount importance for putting situation-specific public health measures in place.

2.4. Impacts of Variants

2.4.1. Impact on diagnoses tools

Sequencing data can be used to understand whether current diagnostic tests are still fit for purpose, as well as to support the development of new diagnostics.

Indeed, the effects of the emergence of new variants include the fact that virus mutations can potentially reduce the accuracy of diagnostic tests.

Therefore, regular analysis of sequence data can allow researchers to identify any mutations of particular concern, which can then be studied to see if they have have an impact on the function of the test. Specifically, if a test does not perform as expected, for instance by continuously returning false-negative results, samples could then be sequenced to see if they contain a mutation responsible for the test failure [6].

2.4.2. Impact on public health measures

Newly discovered variants are often more transmissible and lead to higher infection rates and higher levels of hospitalizations, increasing pressure on health care systems.

Although current public health measures, such as social distancing, sanitizing public places, quarantine and the use of masks, are still effective, adjustments are needed to counteract the increased transmissibility.

There is also concern about the emergence of variants leading to so-called immune escape, with the possibility of reinfection in a shorter time frame, resulting in repeated infections [6].

2.4.3. Impact on vaccines

One of the major fears associated with the emergence and spread of variants concerns the impact of mutations on vaccine efficacy.
Indeed, there is a possibility that the mutations they possess allow the virus to escape the immunity conferred by vaccination [6].

For example, there is growing evidence to suggest that the B.1.351 variant possesses mutations that allow it to escape the immunity conferred by previous infection and some vaccines [11]. The strongest and most conclusive evidence to date suggests that Oxford-AstraZeneca's COVID-19 vaccine offers only minimal protection against B.1.351 infection [25]. Again, genomic analysis can enable timely updating of vaccines to make them effective even against the most dangerous variants.

3 | Tool Description

Variant Hunter is a highly flexible and user-friendly tool for monitoring the evolution of SARS-CoV-2 at continental, national, and regional level. The tool is based on simple yet effective statistics and visual representations (charts and heat-maps) to track viral evolution and highlight the increase or decrease in prevalence of a given amino acid change or group of amino acid changes.

Variant Hunter mainly supports two types of analysis, namely lineage-independent and lineage-specific analysis. Both are based on the same idea: to examine all relevant amino acid mutations detected in the sequences over a four-week period. Specifically, for each week, the frequency of each mutation is considered. The latter is computed as m_count/tot_count (the blue dots in Figure 3.1), where m_count is the number of sequences affected by the considered mutation and tot_count is the total number of sequences collected in that week . Next, a linear model is fitted on the four data points (the red line in Figure 3.1).

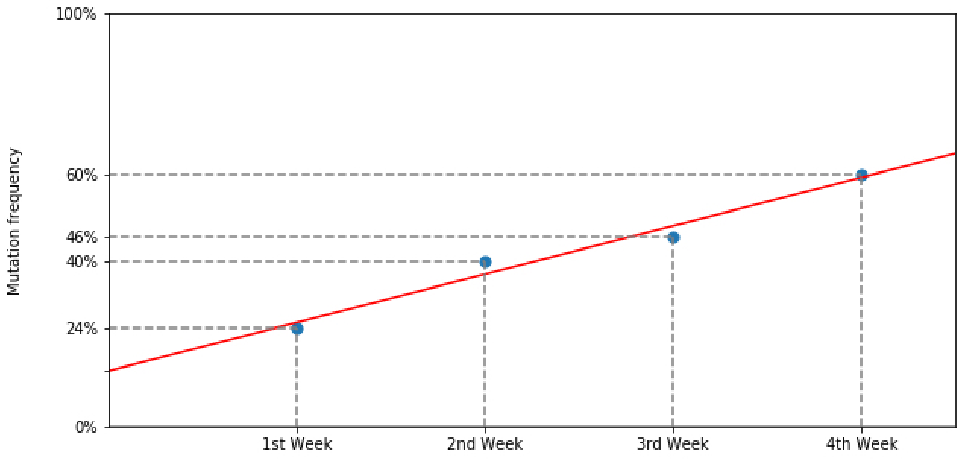


Figure 3.1: Mutation frequency plot.

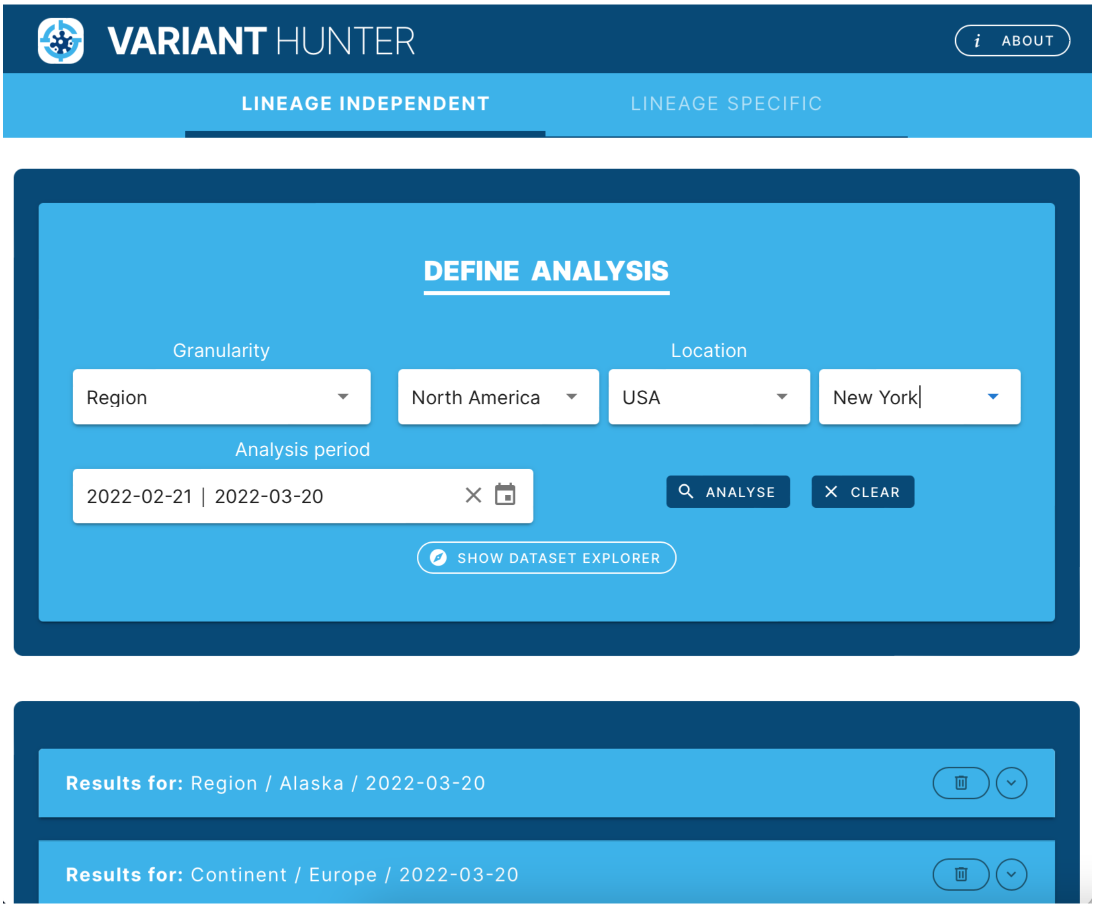
Of particular interest is the slope of the regression line, which represents the growth rate of a mutation (i.e., how fast a mutation is growing in percentage). The slope is very informative and intuitive, as it does not require statistical knowledge: positive slopes indicate an increasing trend, while negative values of the slope indicate a decreasing trend in the target population. Moreover, Chi squared tests are computed to provide information about the significance of the change in the frequency.

The results are summarized in a table of mutations, reporting their prevalence over the 4 weeks, their slope, and the computed p-values. In addition, different visualizations are provided to convey relevant patterns at a glance.

In the following, the main features of the tool are detailed along with some interesting use cases.

3.1. Lineage Independent Analyses

As previously mentioned, the Lineage Independent feature supports the analysis of the mutation data in a specific time period, and for a user selected geographical area. In the example of Figure 3.2, *New York* and the four weeks ending on *March 20th, 2022* are considered.



Granularity: Region

Location: North America | USA | New York

Analysis period: 2022-02-21 | 2022-03-20

Results for: Region / Alaska / 2022-03-20

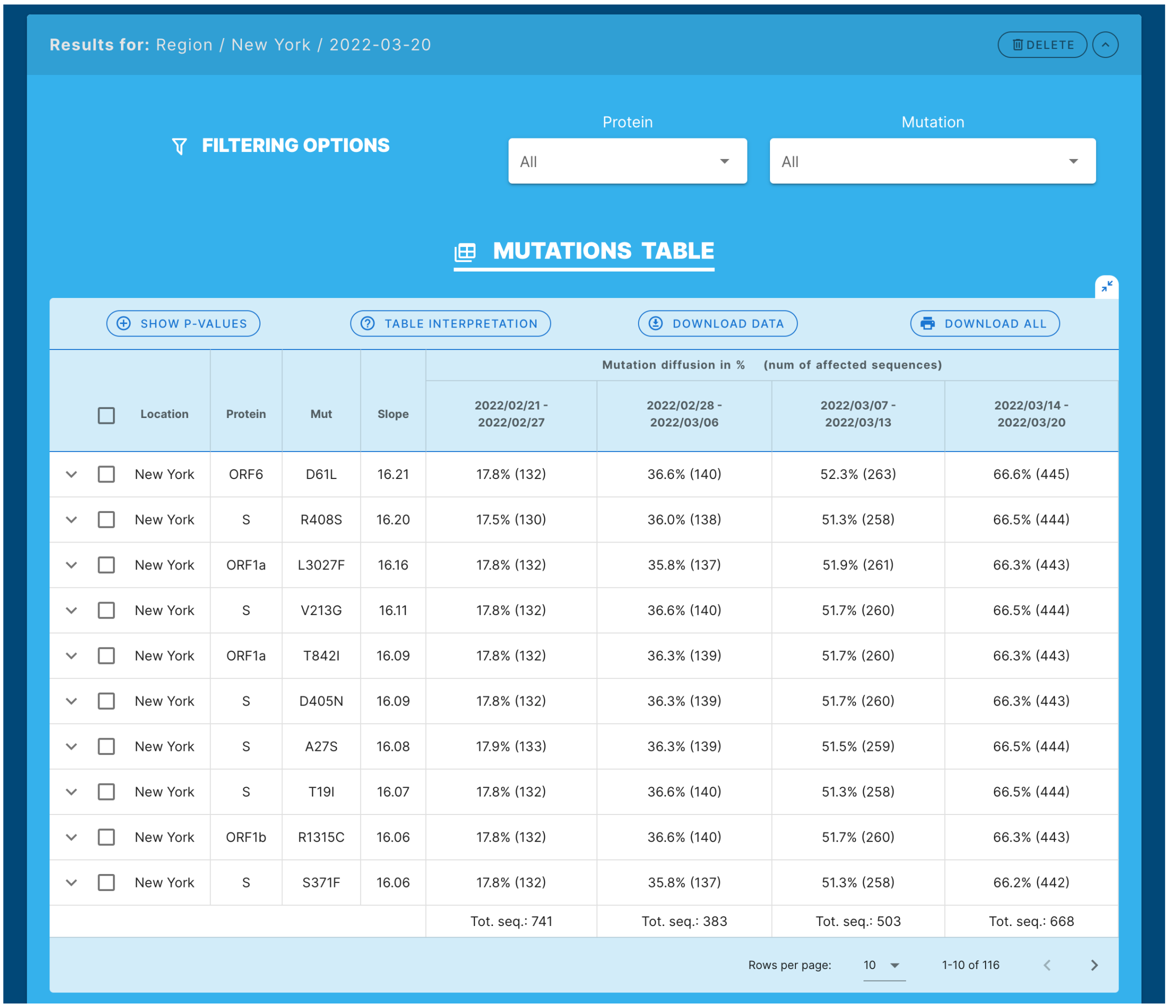
Results for: Continent / Europe / 2022-03-20

Figure 3.2: Lineage Independent Analysis: analysis definition interface.

3.1.1. Mutations table

The results of the analysis are organized in a table like the one shown in Figure 3.3:

- each line represents a mutation of interest;
- the slope represents the rate at which the mutation frequency increased or decreased over the selected period;
- the next four columns report the frequency of the mutation in each of the four weeks



Results for: Region / New York / 2022-03-20

Protein: All

Mutation: All

	□	Location	Protein	Mut	Slope	Mutation diffusion in % (num of affected sequences)			
						2022/02/21 - 2022/02/27	2022/02/28 - 2022/03/06	2022/03/07 - 2022/03/13	2022/03/14 - 2022/03/20
⌄	□	New York	ORF6	D61L	16.21	17.8% (132)	36.6% (140)	52.3% (263)	66.6% (445)
⌄	□	New York	S	R408S	16.20	17.5% (130)	36.0% (138)	51.3% (258)	66.5% (444)
⌄	□	New York	ORF1a	L3027F	16.16	17.8% (132)	35.8% (137)	51.9% (261)	66.3% (443)
⌄	□	New York	S	V213G	16.11	17.8% (132)	36.6% (140)	51.7% (260)	66.5% (444)
⌄	□	New York	ORF1a	T842I	16.09	17.8% (132)	36.3% (139)	51.7% (260)	66.3% (443)
⌄	□	New York	S	D405N	16.09	17.8% (132)	36.3% (139)	51.7% (260)	66.3% (443)
⌄	□	New York	S	A27S	16.08	17.9% (133)	36.3% (139)	51.5% (259)	66.5% (444)
⌄	□	New York	S	T19I	16.07	17.8% (132)	36.6% (140)	51.3% (258)	66.5% (444)
⌄	□	New York	ORF1b	R1315C	16.06	17.8% (132)	36.6% (140)	51.7% (260)	66.3% (443)
⌄	□	New York	S	S371F	16.06	17.8% (132)	35.8% (137)	51.3% (258)	66.2% (442)
						Tot. seq.: 741	Tot. seq.: 383	Tot. seq.: 503	Tot. seq.: 668

Rows per page: 10 1-10 of 116

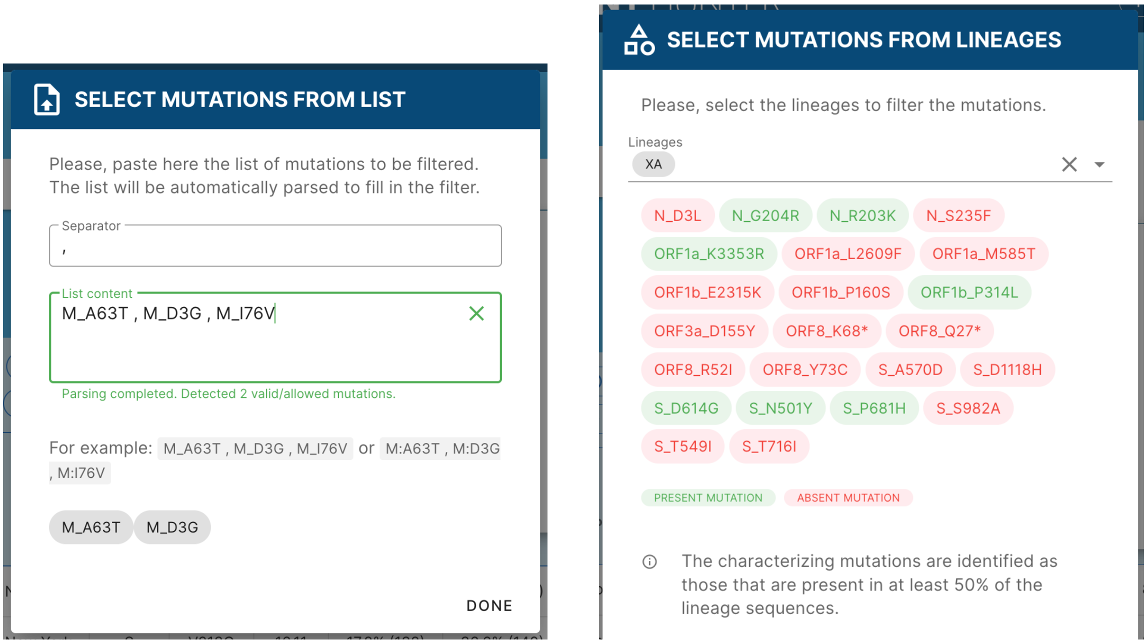
Figure 3.3: Lineage Independent Analysis: Mutation Table interface.

- through a dedicated switch, it is also possible to display three additional columns containing the p-values. These are computed using a Chi-square test of independence of variables in a contingency table, and represents:
  - *P-value with mut*: if the population «with mutation» is growing differently compared to everything;

- *P-value without mut*: if the population «without mutation» is growing differently compared to everything;
- *P-value comparative*: if the population «with mutation» is growing differently compared to the population «without mutation».

The tool proves to be particularly flexible, as it offers options for sorting rows by considering the values in one or more columns, as well as advanced filtering and selection features. As a matter of fact, the user can focus the report on a specific protein or specific mutations. These can be selected in various ways:

- from a drop-down menu;
- from a manually written list, through the interface shown in Figure 3.4a;
- by selecting one or more lineages, from the drop-down menu reported in Figure 3.4b. In this case only mutations characterizing the lineage(s) (i.e., those present in at least half of the lineage sequences) are considered.



SELECT MUTATIONS FROM LIST

Please, paste here the list of mutations to be filtered.
The list will be automatically parsed to fill in the filter.

Separator: ,

List content: M_A63T , M_D3G , M_I76V

Parsing completed. Detected 2 valid/allowed mutations.

For example: M_A63T , M_D3G , M_I76V or M:A63T , M:D3G , M:I76V

M_A63T M_D3G

DONE

SELECT MUTATIONS FROM LINEAGES

Please, select the lineages to filter the mutations.

Lineages: XA

N_D3L N_G204R N_R203K N_S235F
ORF1a_K3353R ORF1a_L2609F ORF1a_M585T
ORF1b_E2315K ORF1b_P160S ORF1b_P314L
ORF3a_D155Y ORF8_K68* ORF8_Q27*
ORF8_R52I ORF8_Y73C S_A570D S_D1118H
S_D614G S_N501Y S_P681H S_S982A
S_T549I S_T716I

PRESENT MUTATION ABSENT MUTATION

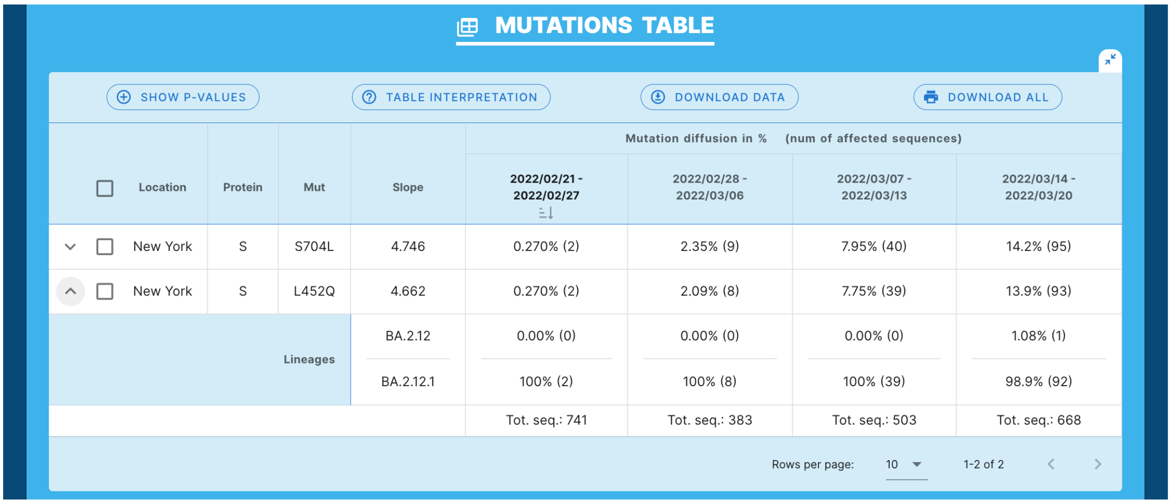
The characterizing mutations are identified as those that are present in at least 50% of the lineage sequences.

(a) Mutations selector from list.

(b) Mutations selector from lineages.

Figure 3.4: Advanced filtering options of Variant Hunter.

Appropriate controls are provided to download the data in various formats.



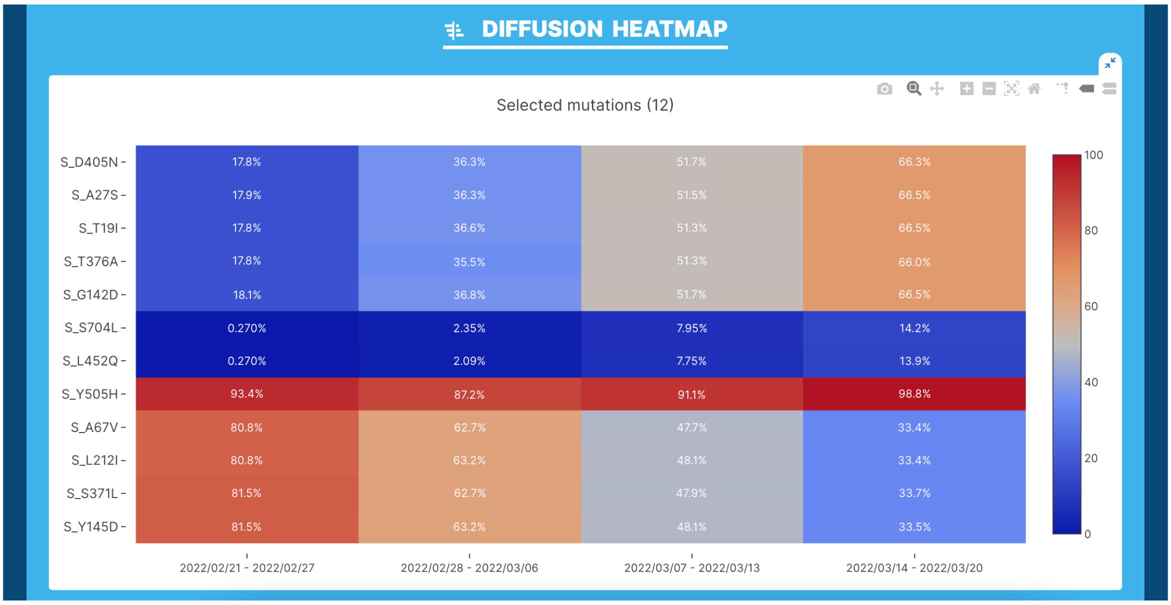
☐	Location	Protein	Mut	Slope	Mutation diffusion in % (num of affected sequences) 2022/02/21 - 2022/02/27	2022/02/28 - 2022/03/06	2022/03/07 - 2022/03/13	2022/03/14 - 2022/03/20
☐	New York	S	S704L	4.746	0.270% (2)	2.35% (9)	7.95% (40)	14.2% (95)
☐	New York	S	L452Q	4.662	0.270% (2)	2.09% (8)	7.75% (39)	13.9% (93)
			Lineages	BA.2.12	0.00% (0)	0.00% (0)	0.00% (0)	1.08% (1)
				BA.2.12.1	100% (2)	100% (8)	100% (39)	98.9% (92)
					Tot. seq.: 741	Tot. seq.: 383	Tot. seq.: 503	Tot. seq.: 668

Figure 3.5: Lineage Independent Analysis: lineage decomposition of a mutation.

This kind of search also provide the possibility to examine the lineage breakdown of the sequences having a given mutation. As shown in Figure 3.5, the expanded section reports the decomposition over the four weeks. In the example, the user can notice that the mutation under consideration is associated almost exclusively with the BA.2.12.1 lineage.

3.1.2. Visualizations

Different visualizations are employed to support the *Lineage Independent Analysis*.



	2022/02/21 - 2022/02/27	2022/02/28 - 2022/03/06	2022/03/07 - 2022/03/13	2022/03/14 - 2022/03/20
S_D405N	17.8%	36.3%	51.7%	66.3%
S_A27S	17.9%	36.3%	51.5%	66.5%
S_T19I	17.8%	36.6%	51.3%	66.5%
S_T376A	17.8%	35.5%	51.3%	66.0%
S_G142D	18.1%	36.8%	51.7%	66.5%
S_S704L	0.270%	2.35%	7.95%	14.2%
S_L452Q	0.270%	2.09%	7.75%	13.9%
S_Y505H	93.4%	87.2%	91.1%	98.8%
S_A67V	80.8%	62.7%	47.7%	33.4%
S_L212I	80.8%	63.2%	48.1%	33.4%
S_S371L	81.5%	62.7%	47.9%	33.7%
S_Y145D	81.5%	63.2%	48.1%	33.5%

Figure 3.6: Lineage Independent Analysis: Diffusion Heatmap.

A **Diffusion Heatmap** allows to observe the increasing and decreasing trends in a quick way. For example in Figure 3.6, a group of mutations (representing *BA.1*) is clearly decreasing, whereas another group is increasing (representing *BA.2*). Other mutations, instead, have less clear trends.

The **Diffusion Trend Chart** (Figure 3.7) allows the user to discover the overtaking point between groups of mutations with opposite trends.

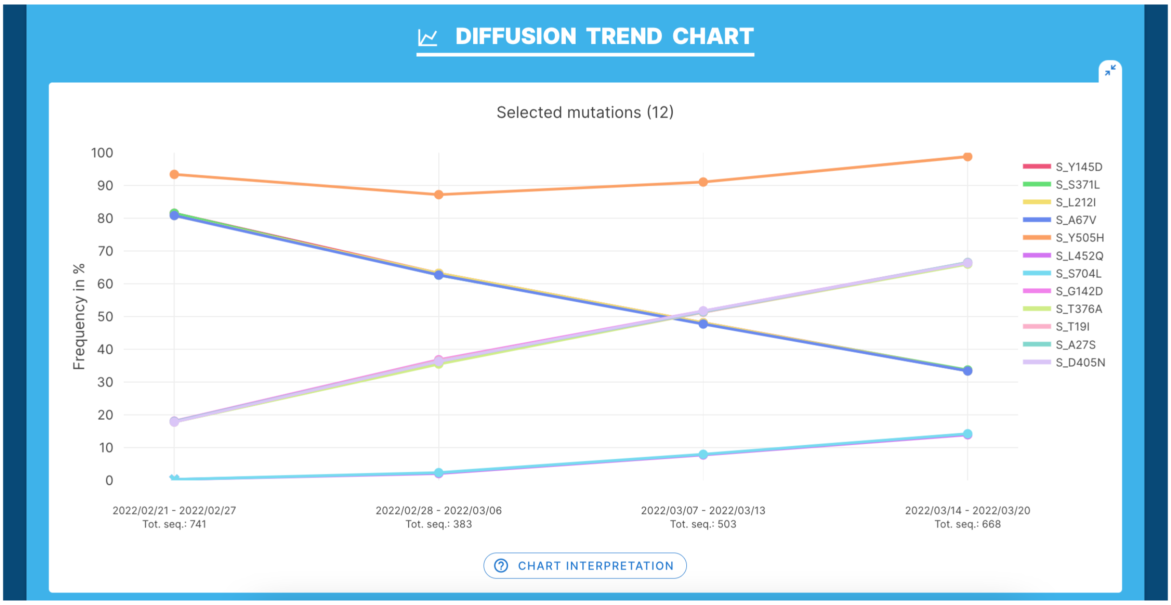


Figure 3.7: Lineage Independent Analysis: diffusion trend chart.

To support the user experience, mutations in the legend on the right can be selected and deselected with a single click. In addition, the graphs provides options to zoom in, move through the axes, or take a screenshot.

In Figure 3.7, for example, it can be seen that at the bottom of the graph, a couple of mutations go from a very low prevalence (about 0%) to almost 15%.

Such an increase can be appreciated more clearly by looking at the **Diffusion Odd Ratio** (Figure 3.8), where the odd ratio is computed by comparing the frequencies of each week with those of the previous week or with those of the first week of the analysis period.

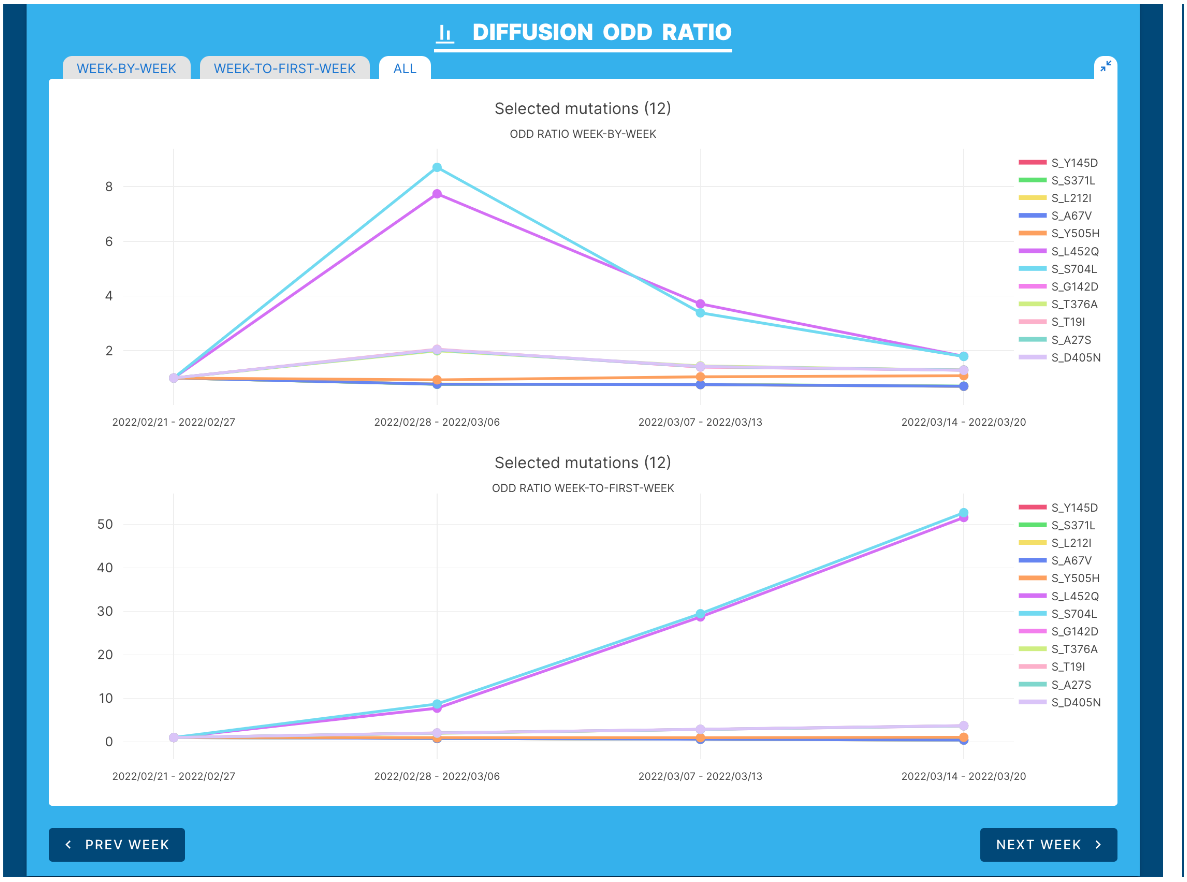


Figure 3.8: Lineage Independent Analysis: diffusion odd ratio chart.

Another interesting feature is the one that allows to navigate through time by replicating the same analysis shifted one week forward or backward. This operation can be easily done by the user through appropriate buttons, shown in Figure 3.8, and is particularly useful to check whether a pattern, identified in the current analysis, is also verified in the subsequent/preceding week.

3.1.3. Use cases

This section provides interesting use cases for the *Lineage Independent* analysis function. Specifically, it is shown how the statistics provided by the tool can be analyzed in order to identify growing mutations and monitor variant trends. Specifically:

- the first use case shows that it is possible to track the trend of a group of mutations of interest over time;
- the second use case shows how it is possible to determine the imminent emergence of a new variant;
- the third and last use case, on the other hand, focuses on a later, more advanced stage of spread and deals with the case when one variant becomes predominant over another;

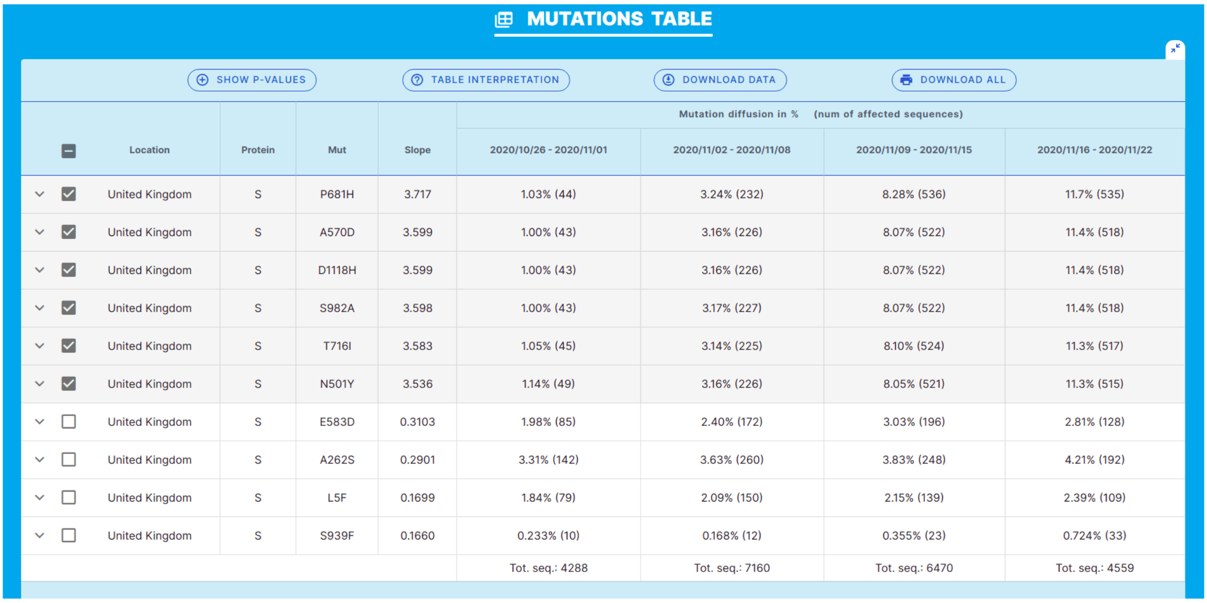
All three use cases are based on sequencing data made available by Nextstrain as of 19/04/2022.

Emergence and spread of Alpha variant in the UK

As a first possible use case, let us consider the following in which the *Lineage Independent* feature allows us to **detect the appearance and spread of a new variant**, namely the Alpha one in this case.

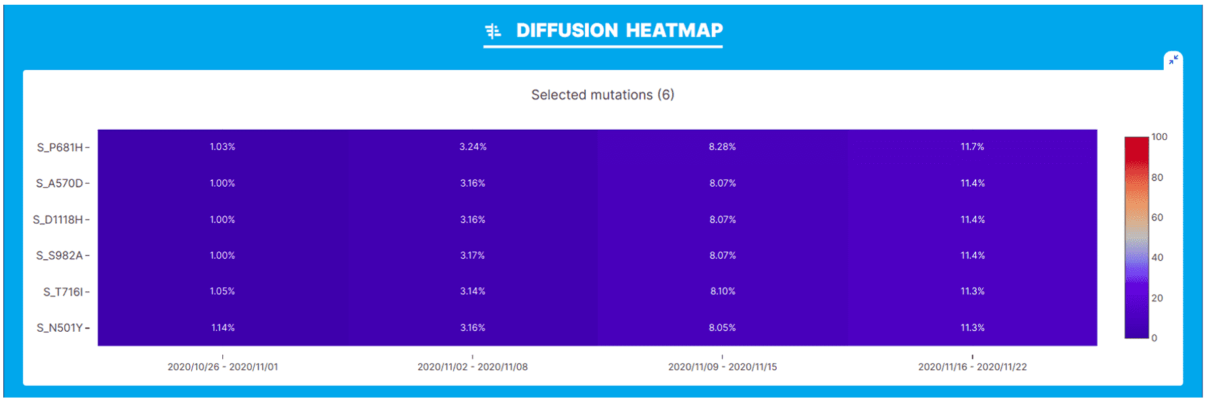
The following is a summary of the main results of the analysis performed over the four-week interval from 26/10/2020 to 16/11/2020 in the United Kingdom. This time frame corresponds to the emergence and spread of the Alpha variant in the UK.
Nextstrain data available as of 19/04/2022 were used for this experiment.

As illustrated by the *Mutations Table* in Figure 3.9a and the *Diffusion Heatmap* (Figure 3.9b), 6 amino acid changes in the Spike glycoprotein stand out due to the significant increase in frequency (from 1% to 11%).



	Location	Protein	Mut	Slope	Mutation diffusion in % (num of affected sequences)			
▬					2020/10/26 - 2020/11/01	2020/11/02 - 2020/11/08	2020/11/09 - 2020/11/15	2020/11/16 - 2020/11/22
☑	United Kingdom	S	P681H	3.717	1.03% (44)	3.24% (232)	8.28% (536)	11.7% (535)
☑	United Kingdom	S	A570D	3.599	1.00% (43)	3.16% (226)	8.07% (522)	11.4% (518)
☑	United Kingdom	S	D1118H	3.599	1.00% (43)	3.16% (226)	8.07% (522)	11.4% (518)
☑	United Kingdom	S	S982A	3.598	1.00% (43)	3.17% (227)	8.07% (522)	11.4% (518)
☑	United Kingdom	S	T716I	3.583	1.05% (45)	3.14% (225)	8.10% (524)	11.3% (517)
☑	United Kingdom	S	N501Y	3.536	1.14% (49)	3.16% (226)	8.05% (521)	11.3% (515)
☐	United Kingdom	S	E583D	0.3103	1.98% (85)	2.40% (172)	3.03% (196)	2.81% (128)
☐	United Kingdom	S	A262S	0.2901	3.31% (142)	3.63% (260)	3.83% (248)	4.21% (192)
☐	United Kingdom	S	L5F	0.1699	1.84% (79)	2.09% (150)	2.15% (139)	2.39% (109)
☐	United Kingdom	S	S939F	0.1660	0.233% (10)	0.168% (12)	0.355% (23)	0.724% (33)
					Tot. seq.: 4288	Tot. seq.: 7160	Tot. seq.: 6470	Tot. seq.: 4559

(a) Mutation Table.



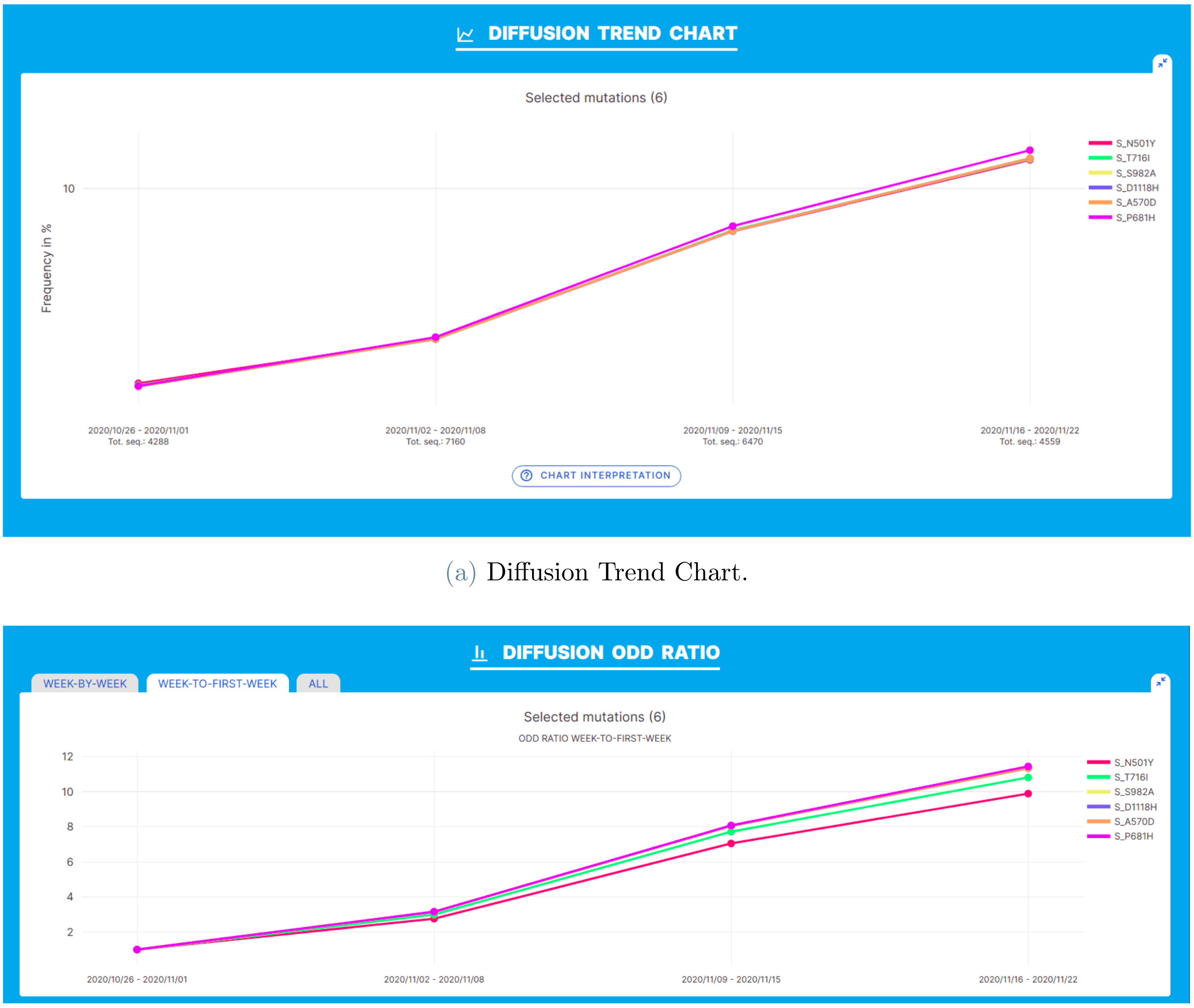
(b) Diffusion Heatmap.

Figure 3.9: Mutation Table and heatmap of the lineage independent analysis for the period 26/10/2022 - 16/11/2020 in the UK. Data from Nextstrain dated 19/04/2022.

These amino acid changes represent the entire set of amino acid modifications in Spike of the Alpha variant of SARS-CoV-2. With Variant Hunter, it is possible to analyze their trends in detail by selecting the rows of interest from the table.

In this way, the subsequent graphs, which by default represent the 5 mutations with the greatest growth and the 5 with the greatest decrease, will instead show the data from the selected rows.

Similar patterns of frequency increase of the six changes can also be observed from the *Diffusion Trend Chart* and the *Diffusion Odd Ratio* plots in Figure 3.10.



(a) Diffusion Trend Chart.

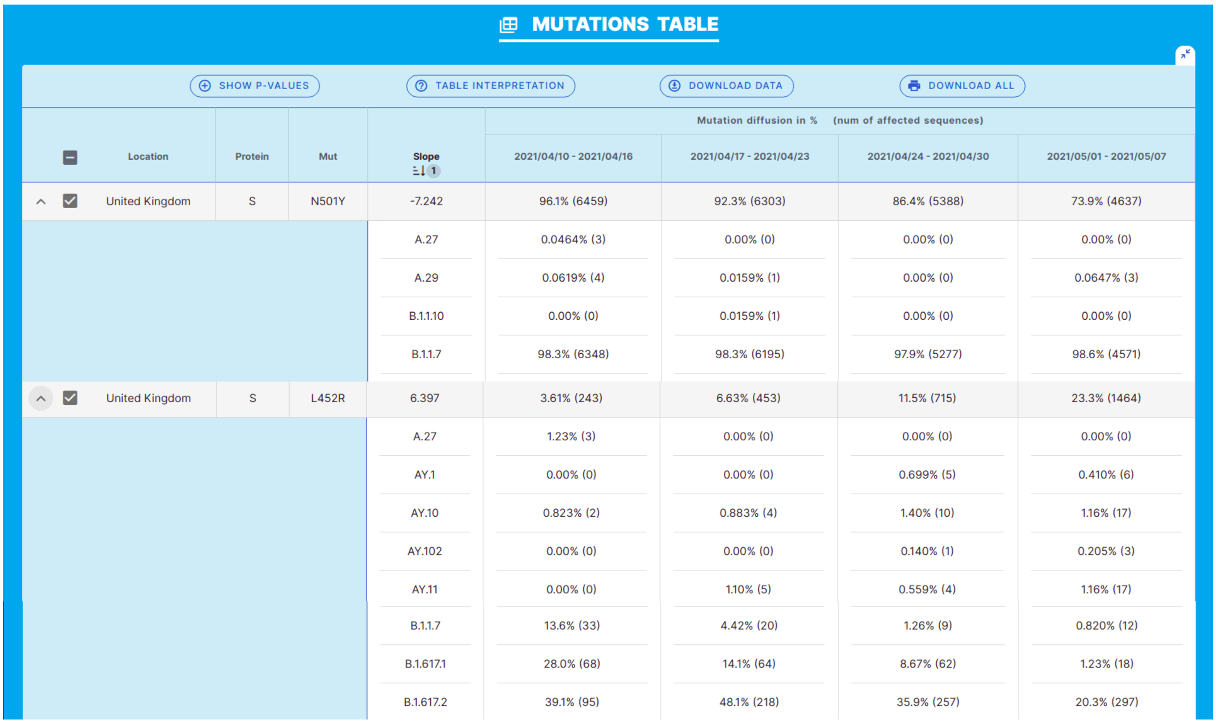
(b) Diffusion Odd Ratio.

Figure 3.10: Diffusion trend chart and odd ratio of the lineage independent analysis for the period 26/10/2022 - 16/11/2020 in the UK. Data from Nextstrain dated 19/04/2022.

Early spread of the Delta variant in the UK

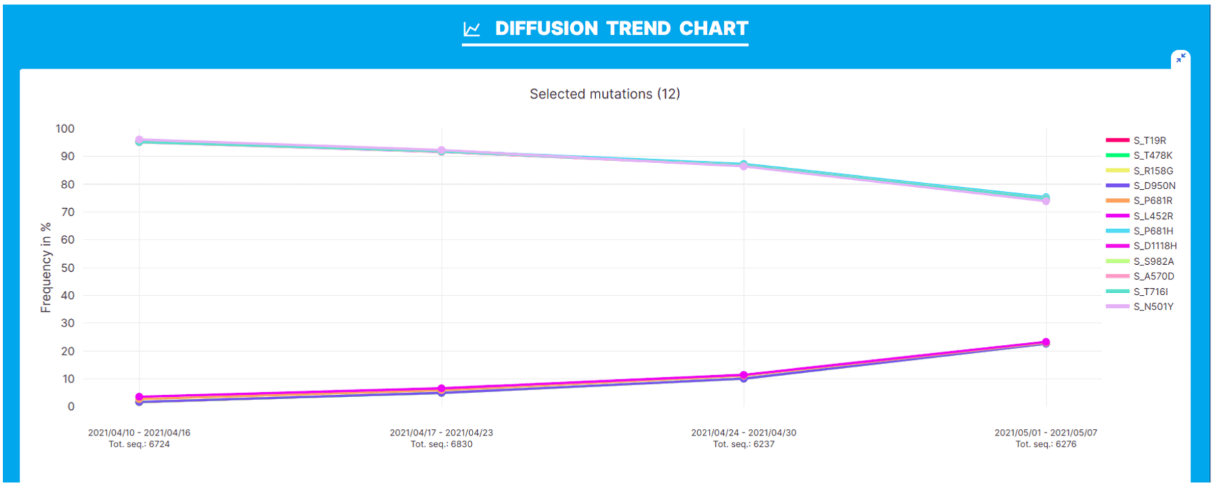
Another possible usage of the *Lineage Independent* analysis feature is the **early discovery of a new emerging variant**. For example, leveraging Nextstrain data available as of 04/19/2022, an analysis can be performed on the time interval from 10/04/2021 to 07/05/2021 in the United Kingdom. This four-weeks range corresponds to the first spread of the Delta variant in the UK.

The Mutation Table (Figure 3.11a) illustrates the relative prevalence of `S:N501Y` and `S:L452R`, two amino acid changes characteristic of Alpha and Delta, respectively, in different sub-lineages of SARS-CoV-2. While `S:N501Y` is mostly associated with Alpha-related lineages, S:L452R is observed almost exclusively in Delta lineages/sub-lineages.



	Location	Protein	Mut	Slope	Mutation diffusion in % (num of affected sequences)			
☐					2021/04/10 - 2021/04/16	2021/04/17 - 2021/04/23	2021/04/24 - 2021/04/30	2021/05/01 - 2021/05/07
☑	United Kingdom	S	N501Y	-7.242	96.1% (6459)	92.3% (6303)	86.4% (5388)	73.9% (4637)
				A.27	0.0464% (3)	0.00% (0)	0.00% (0)	0.00% (0)
				A.29	0.0619% (4)	0.0159% (1)	0.00% (0)	0.0647% (3)
				B.1.1.10	0.00% (0)	0.0159% (1)	0.00% (0)	0.00% (0)
				B.1.1.7	98.3% (6348)	98.3% (6195)	97.9% (5277)	98.6% (4571)
☑	United Kingdom	S	L452R	6.397	3.61% (243)	6.63% (453)	11.5% (715)	23.3% (1464)
				A.27	1.23% (3)	0.00% (0)	0.00% (0)	0.00% (0)
				AY.1	0.00% (0)	0.00% (0)	0.699% (5)	0.410% (6)
				AY.10	0.823% (2)	0.883% (4)	1.40% (10)	1.16% (17)
				AY.102	0.00% (0)	0.00% (0)	0.140% (1)	0.205% (3)
				AY.11	0.00% (0)	1.10% (5)	0.559% (4)	1.16% (17)
				B.1.1.7	13.6% (33)	4.42% (20)	1.26% (9)	0.820% (12)
				B.1.617.1	28.0% (68)	14.1% (64)	8.67% (62)	1.23% (18)
				B.1.617.2	39.1% (95)	48.1% (218)	35.9% (257)	20.3% (297)

(a) Expanded Mutation Table.



(b) Diffusion Trend Chart.

Figure 3.11: Mutation Table and diffusion trend chart of the lineage independent analysis for the period 10/04/2021 - 07/05/2021 in the UK. Data from Nextstrain dated 19/04/2022.

A clear increase in changes associated with the Delta variant (such as S:L452R) and a concomitant decrease in changes characteristic of the Alpha variant (such as S:N501Y) can be observed from the diffusion trend chart (Figure 3.11b).

Spread of the Omicron variant and displacement of Delta in the Europe and North America

Finally, the *Lineage Independent* search can be used also **to observe when one variant becomes predominant over another**. For this use case, the period from 05/12/2021 to 01/01/2022 in Europe and North America was examined. This interval coincides with the rapid spread of the Omicron variant of SARS-CoV-2 worldwide. The objective here is to investigate the prevalence of amino acid changes in the Spike glycoprotein.

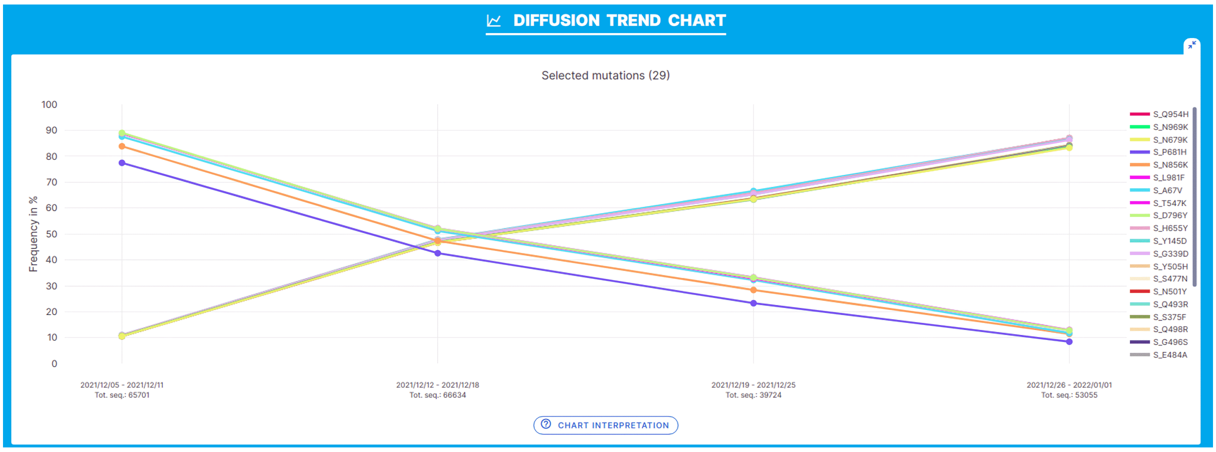
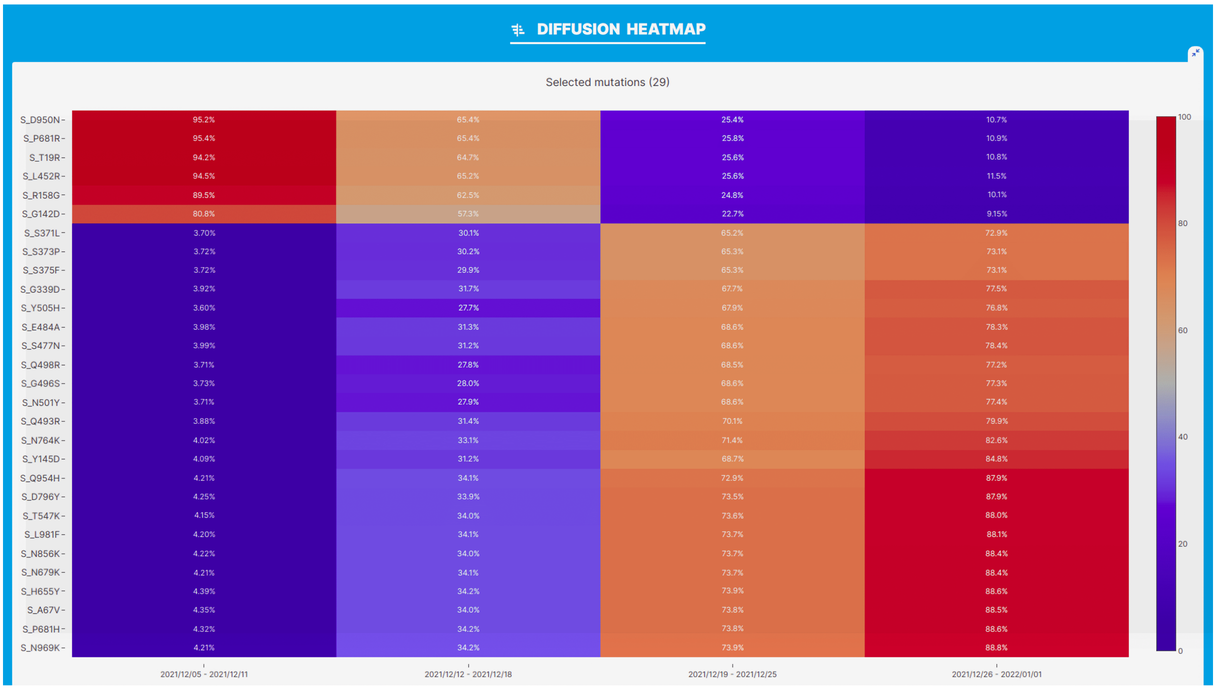


Figure 3.12: Diffusion Trend Chart of the lineage independent analysis for the period 05/12/2021 - 01/01/2022 in Europe. Data from Nextstrain dated 19/04/2022.

In both Europe and North America, there is a marked decrease in the prevalence of 6 amino acid changes characteristic of the Delta variant (`L452R`, `P681R`, `D950N`, `R158G`, `S478K`, `T19R`) and a concomitant increase in 23 amino acid changes associated with the Omicron variant of SARS-CoV-2.

Equivalent patterns are derived from both the Diffusion Heatmap (Figure 3.13) and Diffusion Trend Chart (Figure 3.12).



DIFFUSION HEATMAP

Selected mutations (29)

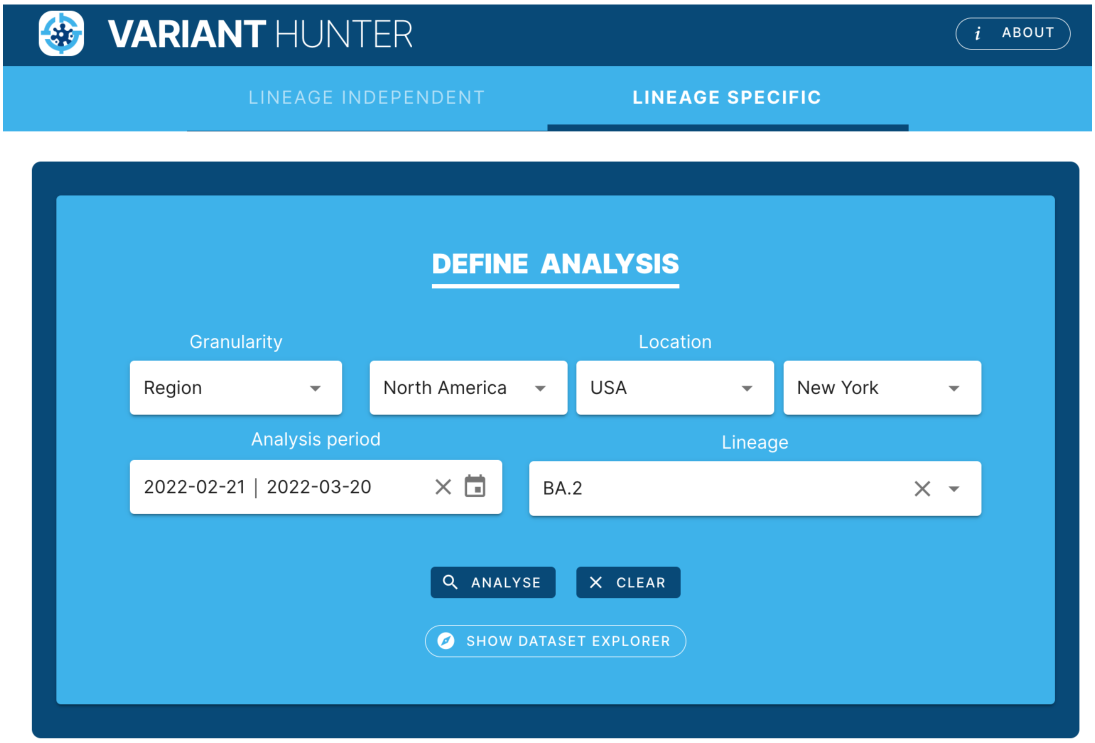
	2021/12/05 - 2021/12/11	2021/12/12 - 2021/12/18	2021/12/19 - 2021/12/25	2021/12/26 - 2022/01/01
S_D950N	95.2%	65.4%	25.4%	10.7%
S_P681R	95.4%	65.4%	25.8%	10.9%
S_T19R	94.2%	64.7%	25.6%	10.8%
S_L452R	94.5%	65.2%	25.6%	11.5%
S_R158G	89.5%	62.5%	24.8%	10.1%
S_G142D	80.8%	57.3%	22.7%	9.15%
S_S371L	3.70%	30.1%	65.2%	72.9%
S_S373P	3.72%	30.2%	65.3%	73.1%
S_S375F	3.72%	29.9%	65.3%	73.1%
S_G339D	3.92%	31.7%	67.7%	77.5%
S_Y505H	3.60%	27.7%	67.9%	76.8%
S_E484A	3.98%	31.3%	68.6%	78.3%
S_S477N	3.99%	31.2%	68.6%	78.4%
S_Q498R	3.71%	27.8%	68.5%	77.2%
S_G496S	3.73%	28.0%	68.6%	77.3%
S_N501Y	3.71%	27.9%	68.6%	77.4%
S_Q493R	3.88%	31.4%	70.1%	79.9%
S_N764K	4.02%	33.1%	71.4%	82.6%
S_Y145D	4.09%	31.2%	68.7%	84.8%
S_Q954H	4.21%	34.1%	72.9%	87.9%
S_D796Y	4.25%	33.9%	73.5%	87.9%
S_T547K	4.15%	34.0%	73.6%	88.0%
S_L981F	4.20%	34.1%	73.7%	88.1%
S_N856K	4.22%	34.0%	73.7%	88.4%
S_N679K	4.21%	34.1%	73.7%	88.4%
S_H655Y	4.39%	34.2%	73.9%	88.6%
S_A67V	4.35%	34.0%	73.8%	88.5%
S_P681H	4.32%	34.2%	73.8%	88.6%
S_N969K	4.21%	34.2%	73.9%	88.8%

Figure 3.13: Diffusion Heatmap of the lineage independent analysis for the period 05/12/2021 - 01/01/2022 in North America. Data from Nextstrain dated 19/04/2022.

The higher uniformity of graphs depicting mutation trends in Europe compared to North America may indicate greater accuracy of sequencing processes.

3.2. Lineage Specific Analyses

As previously introduced, the Lineage Specific function enables the analysis of amino acid substitutions over a specific time period and for a user-selected location and lineage. For instance, considering the example in Figure 3.14, mutations associated with lineage *BA.2* in *New York* are extracted.



Granularity: Region

Location: North America, USA, New York

Analysis period: 2022-02-21 | 2022-03-20

Lineage: BA.2

Figure 3.14: Lineage Specific Analysis: analysis definition interface.

3.2.1. Mutations table

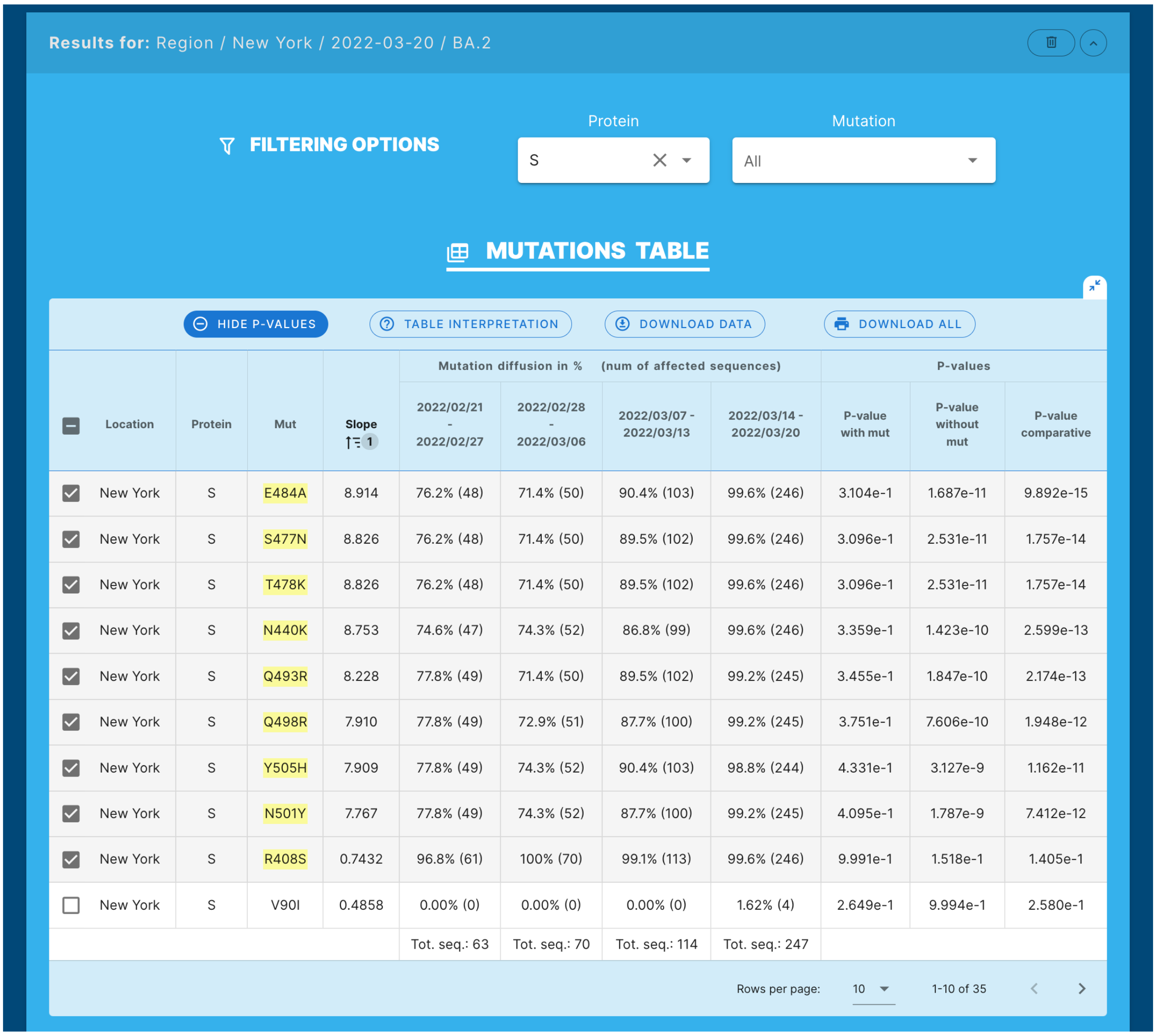
Also for this feature, the results of the analysis are arranged in a table as the one in Figure 3.15. It assumes the same interpretation given in Section 3.1:

- each line represents a mutation of interest;
- the slope represents the rate at which the mutation frequency increased or decreased over the selected period;
- the next four columns report the frequency of the mutation in each of the four weeks.

Clearly, in this case, the counts given for each mutation only account for the sequences of the selected lineage found in each week.

As before, via an appropriate switch it is possible to show the p-values. They are computed using a Chi-square test of independence of variables in a contingency table, and represents:

- *P-value with mut*: if the population «with mutation» is growing differently compared to everything;
- *P-value without mut*: if the population «without mutation» is growing differently compared to everything;
- *P-value comparative*: if the population «with mutation» is growing differently compared to the population «without mutation».



Results for: Region / New York / 2022-03-20 / BA.2

FILTERING OPTIONS

Protein: S

Mutation: All

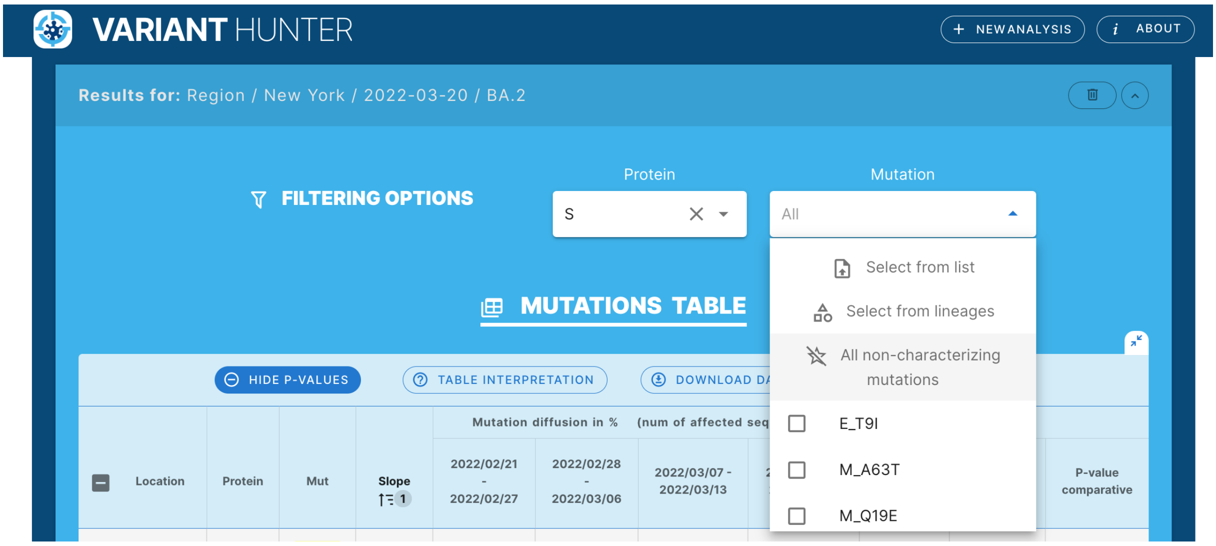
MUTATIONS TABLE

					Mutation diffusion in %	(num of affected sequences)			P-values		
▬	Location	Protein	Mut	Slope ↑ 1	2022/02/21 - 2022/02/27	2022/02/28 - 2022/03/06	2022/03/07 - 2022/03/13	2022/03/14 - 2022/03/20	P-value with mut	P-value without mut	P-value comparative
☑	New York	S	E484A	8.914	76.2% (48)	71.4% (50)	90.4% (103)	99.6% (246)	3.104e-1	1.687e-11	9.892e-15
☑	New York	S	S477N	8.826	76.2% (48)	71.4% (50)	89.5% (102)	99.6% (246)	3.096e-1	2.531e-11	1.757e-14
☑	New York	S	T478K	8.826	76.2% (48)	71.4% (50)	89.5% (102)	99.6% (246)	3.096e-1	2.531e-11	1.757e-14
☑	New York	S	N440K	8.753	74.6% (47)	74.3% (52)	86.8% (99)	99.6% (246)	3.359e-1	1.423e-10	2.599e-13
☑	New York	S	Q493R	8.228	77.8% (49)	71.4% (50)	89.5% (102)	99.2% (245)	3.455e-1	1.847e-10	2.174e-13
☑	New York	S	Q498R	7.910	77.8% (49)	72.9% (51)	87.7% (100)	99.2% (245)	3.751e-1	7.606e-10	1.948e-12
☑	New York	S	Y505H	7.909	77.8% (49)	74.3% (52)	90.4% (103)	98.8% (244)	4.331e-1	3.127e-9	1.162e-11
☑	New York	S	N501Y	7.767	77.8% (49)	74.3% (52)	87.7% (100)	99.2% (245)	4.095e-1	1.787e-9	7.412e-12
☑	New York	S	R408S	0.7432	96.8% (61)	100% (70)	99.1% (113)	99.6% (246)	9.991e-1	1.518e-1	1.405e-1
☐	New York	S	V90I	0.4858	0.00% (0)	0.00% (0)	0.00% (0)	1.62% (4)	2.649e-1	9.994e-1	2.580e-1
					Tot. seq.: 63	Tot. seq.: 70	Tot. seq.: 114	Tot. seq.: 247			

Rows per page: 10 1-10 of 35

Figure 3.15: Lineage Specific Analysis: Mutation Table with p-values.

As demonstrated in Figure 3.16, among other filtering options, this type of analysis also offers the possibility to directly select all the non-characterizing mutations for the lineage under consideration (i.e., *BA.2* in this example). Note that in the Mutation Table (Figure 3.15), the characterizing amino acid changes (i.e., those present in at least half of the lineage sequences) are already highlighted in yellow.



Results for: Region / New York / 2022-03-20 / BA.2

Protein: S

Mutation: All

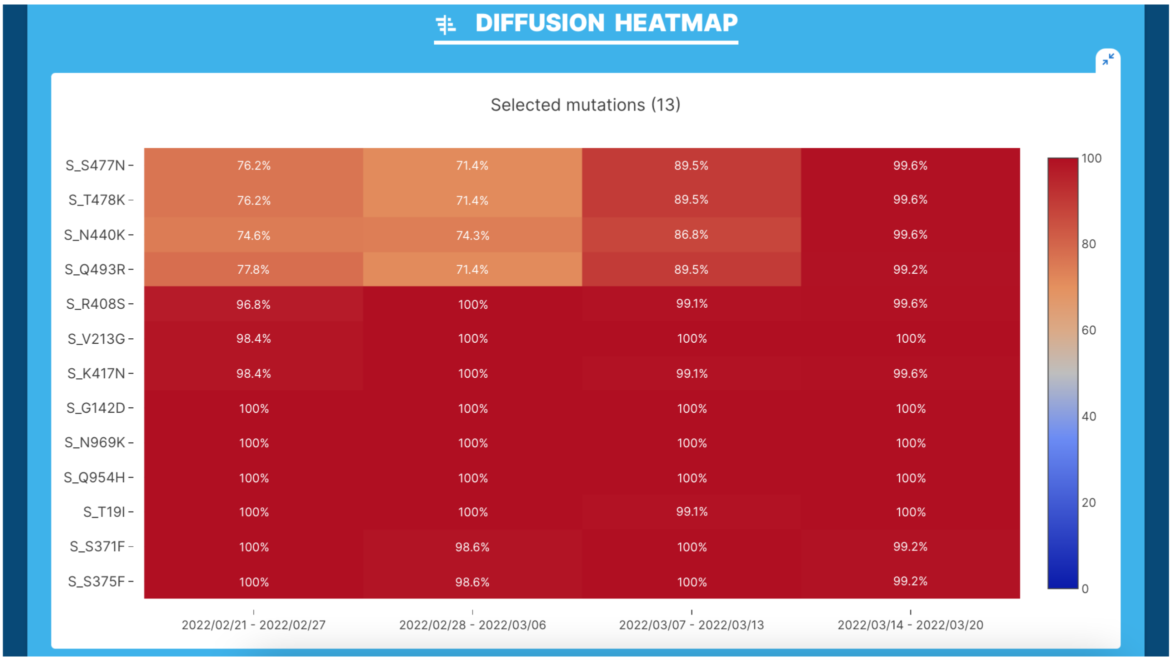
Figure 3.16: Lineage Specific Analysis: advanced mutation filtering options.

3.2.2. Visualizations

Visualizations are used to facilitate the interpretation of data and allow to easily extract meaningful patterns. The types of graphs employed are similar to those described in Section 3.1.

A **Diffusion Heatmap** helps to discover rising and falling trends on the fly.
For instance from the Diffusion Heatmap in Figure 3.17, one can notice that most of the mutations are stable in the considered period.

However, it is also pretty clear that some of them have increased from about 75% to almost 100%.



	2022/02/21 - 2022/02/27	2022/02/28 - 2022/03/06	2022/03/07 - 2022/03/13	2022/03/14 - 2022/03/20
S_S477N	76.2%	71.4%	89.5%	99.6%
S_T478K	76.2%	71.4%	89.5%	99.6%
S_N440K	74.6%	74.3%	86.8%	99.6%
S_Q493R	77.8%	71.4%	89.5%	99.2%
S_R408S	96.8%	100%	99.1%	99.6%
S_V213G	98.4%	100%	100%	100%
S_K417N	98.4%	100%	99.1%	99.6%
S_G142D	100%	100%	100%	100%
S_N969K	100%	100%	100%	100%
S_Q954H	100%	100%	100%	100%
S_T19I	100%	100%	99.1%	100%
S_S371F	100%	98.6%	100%	99.2%
S_S375F	100%	98.6%	100%	99.2%

Figure 3.17: Lineage Specific Analysis: Diffusion Heatmap.

The pattern is also confirmed by the **Diffusion Trend Chart** (Figure 3.18) and the **Diffusion Odd Ratio** plots (Figure 3.19).

The former depicts the evolution of the mutation frequencies over time (in percentage), while the latter exhibits different types of odd ratios:

- *Week-by-week*: a comparison of the frequency for each week against the previous week;
- *Week-to-first*: a comparison of the frequency for each week against the first week of the analysis period under consideration.

Moreover, appropriate labels make it easier to interpret the graph while interacting with it.

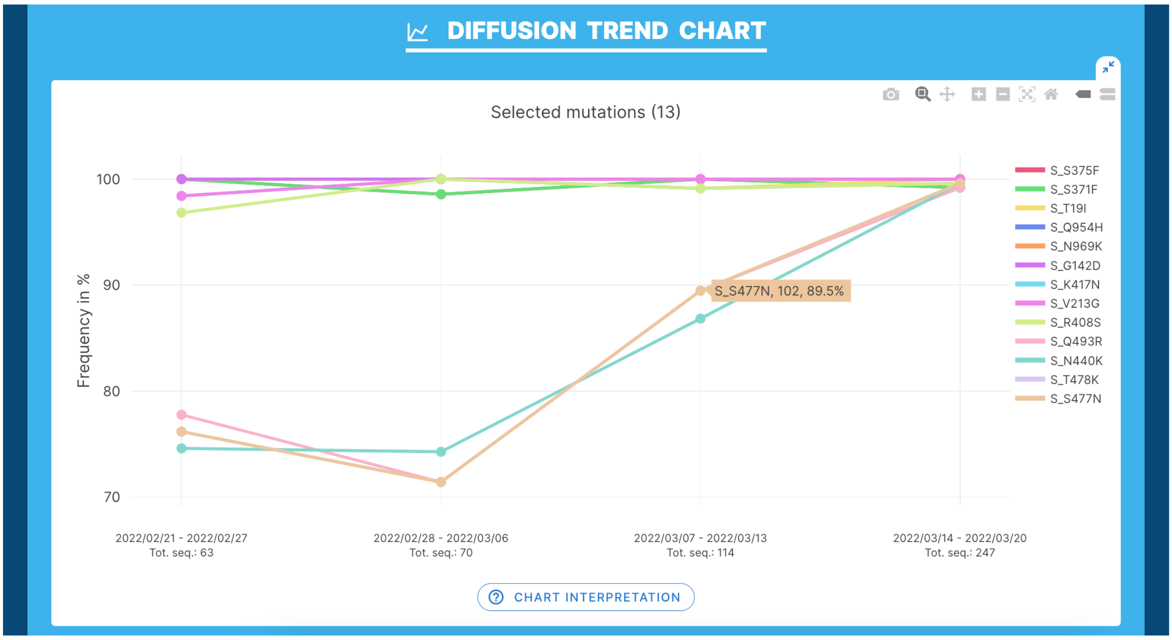
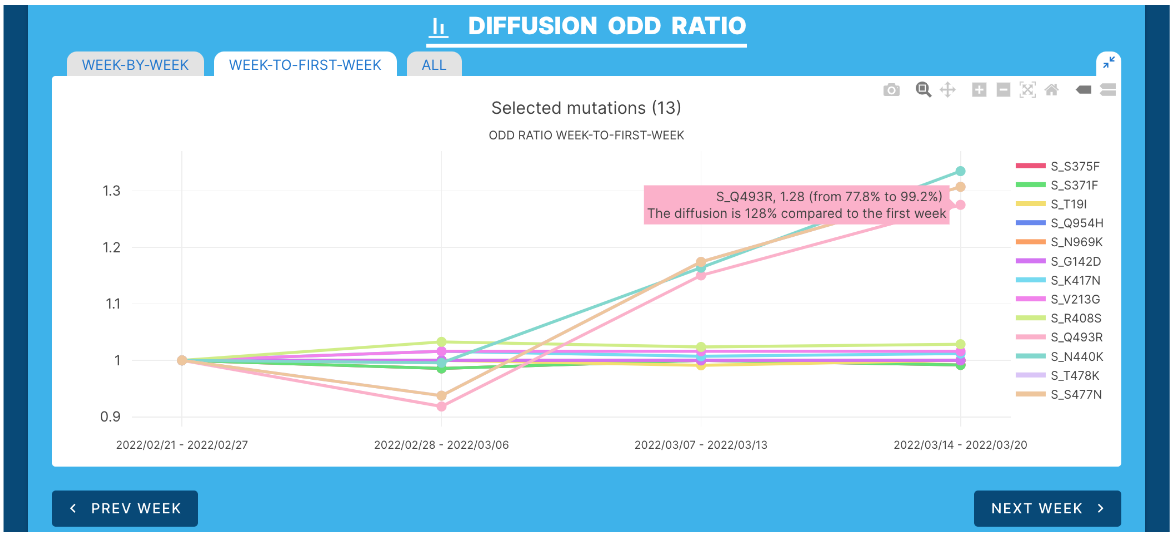


Figure 3.18: Lineage Specific Analysis: diffusion trend chart.



S_Q493R, 1.28 (from 77.8% to 99.2%)
The diffusion is 128% compared to the first week

Figure 3.19: Lineage Specific Analysis: diffusion odd ratio chart.

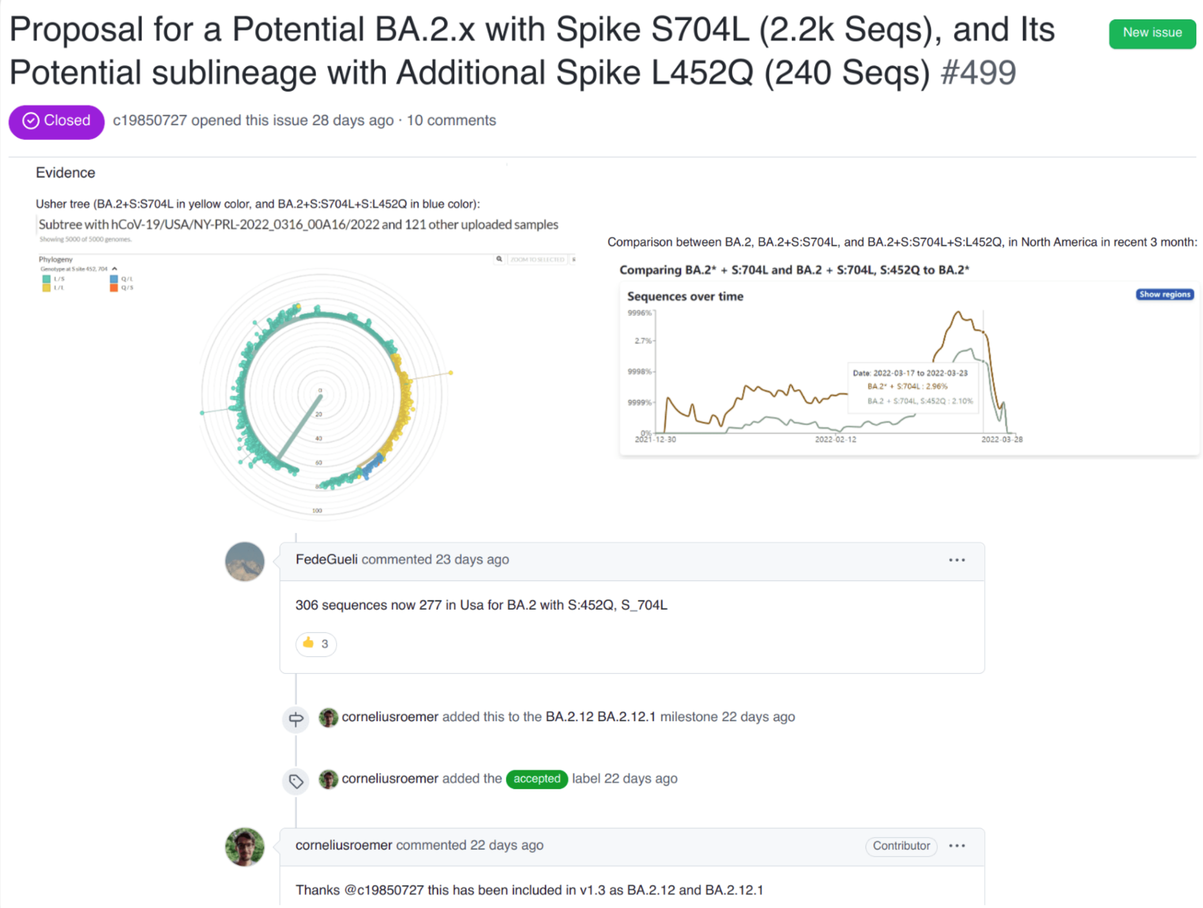
Finally, users are given the ability to shift the analysis period forward or backward by one week. This makes it possible to ascertain whether a given trend endures over time or not.

3.2.3. Use cases

These use cases show how it is possible, through *Lineage Specific* analysis, to verify the occurrence of new mutations within an existing lineage that leads to the definition of a new sub-lineage. In particular, by referring to the submissions on Pango Designation, the repository dedicated to reporting new lineages to be added to the Pangolin [19] scheme, we will attempt to validate such reports through Variant Hunter.

Omicron BA.2.12.2 lineage

In this example we consider a proposal for a potential **sub-lineage of BA.2.12** dated early April 2022.



Proposal for a Potential BA.2.x with Spike S704L (2.2k Seqs), and Its Potential sublineage with Additional Spike L452Q (240 Seqs) #499

Closed c19850727 opened this issue 28 days ago · 10 comments

Evidence

Usher tree (BA.2+S:S704L in yellow color, and BA.2+S:S704L+S:L452Q in blue color):

Subtree with hCoV-19/USA/NY-PRL-2022_0316_00A16/2022 and 121 other uploaded samples

Showing 5000 of 5000 genomes.

Comparison between BA.2, BA.2+S:S704L, and BA.2+S:S704L+S:L452Q, in North America in recent 3 month:

Comparing BA.2* + S:704L and BA.2 + S:704L, S:452Q to BA.2*

FedeGueli commented 23 days ago

306 sequences now 277 in Usa for BA.2 with S:452Q, S_704L

corneliusroemer added this to the BA.2.12 BA.2.12.1 milestone 22 days ago

corneliusroemer added the accepted label 22 days ago

corneliusroemer commented 22 days ago

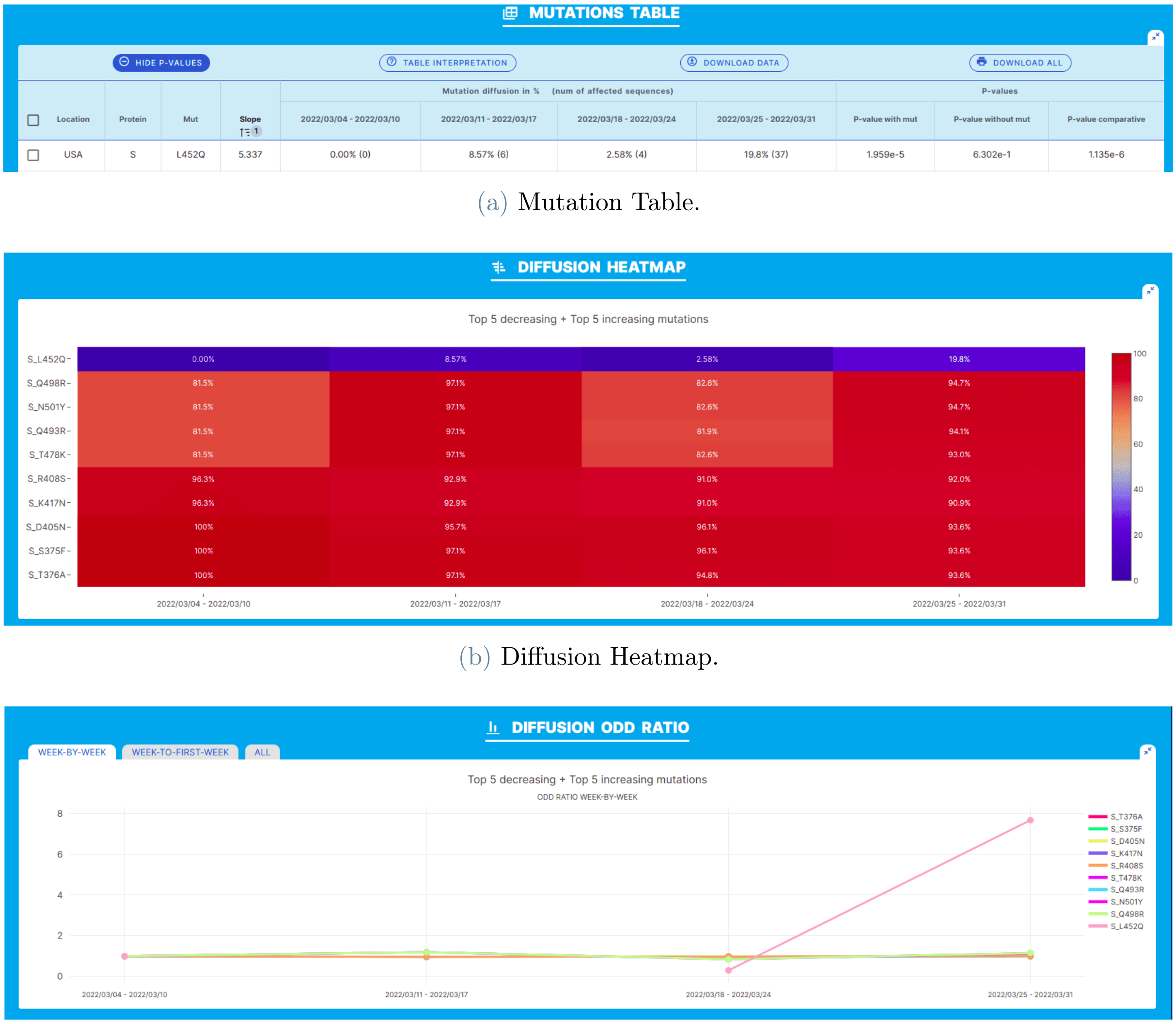
Thanks @c19850727 this has been included in v1.3 as BA.2.12 and BA.2.12.1

Figure 3.20: Issue 499 from Pango designation [8] which led to the definition of the BA.2.12.2 sub-lineage of SARS-Cov-2.

As shown in Figure 3.20, the novel lineage is defined by the amino acid change `L452Q` in the spike glycoprotein, according to the submitter. This report then resulted in the definition of the sub-lineage BA.2.12.2.

To validate this by means of Variant Hunter, it is sufficient to consider a *Lineage Specific* analysis on the BA.2.12 lineage in the geographic location (USA) and interval of time (March 2022) where the novel lineage first emerged.

For the present use case, the Nextstrain data set available as of 04/19/2022 was used, which at the time still lacked the addition of this new classification.



					Mutation diffusion in % (num of affected sequences)				P-values		
☐	Location	Protein	Mut	Slope	2022/03/04 - 2022/03/10	2022/03/11 - 2022/03/17	2022/03/18 - 2022/03/24	2022/03/25 - 2022/03/31	P-value with mut	P-value without mut	P-value comparative
☐	USA	S	L452Q	5.337	0.00% (0)	8.57% (6)	2.58% (4)	19.8% (37)	1.959e-5	6.302e-1	1.135e-6

(a) Mutation Table.

	2022/03/04 - 2022/03/10	2022/03/11 - 2022/03/17	2022/03/18 - 2022/03/24	2022/03/25 - 2022/03/31
S_L452Q	0.00%	8.57%	2.58%	19.8%
S_Q498R	81.5%	97.1%	82.6%	94.7%
S_N501Y	81.5%	97.1%	82.6%	94.7%
S_Q493R	81.5%	97.1%	81.9%	94.1%
S_T478K	81.5%	97.1%	82.6%	93.0%
S_R408S	96.3%	92.9%	91.0%	92.0%
S_K417N	96.3%	92.9%	91.0%	90.9%
S_D405N	100%	95.7%	96.1%	93.6%
S_S375F	100%	97.1%	96.1%	93.6%
S_T376A	100%	97.1%	94.8%	93.6%

(b) Diffusion Heatmap.

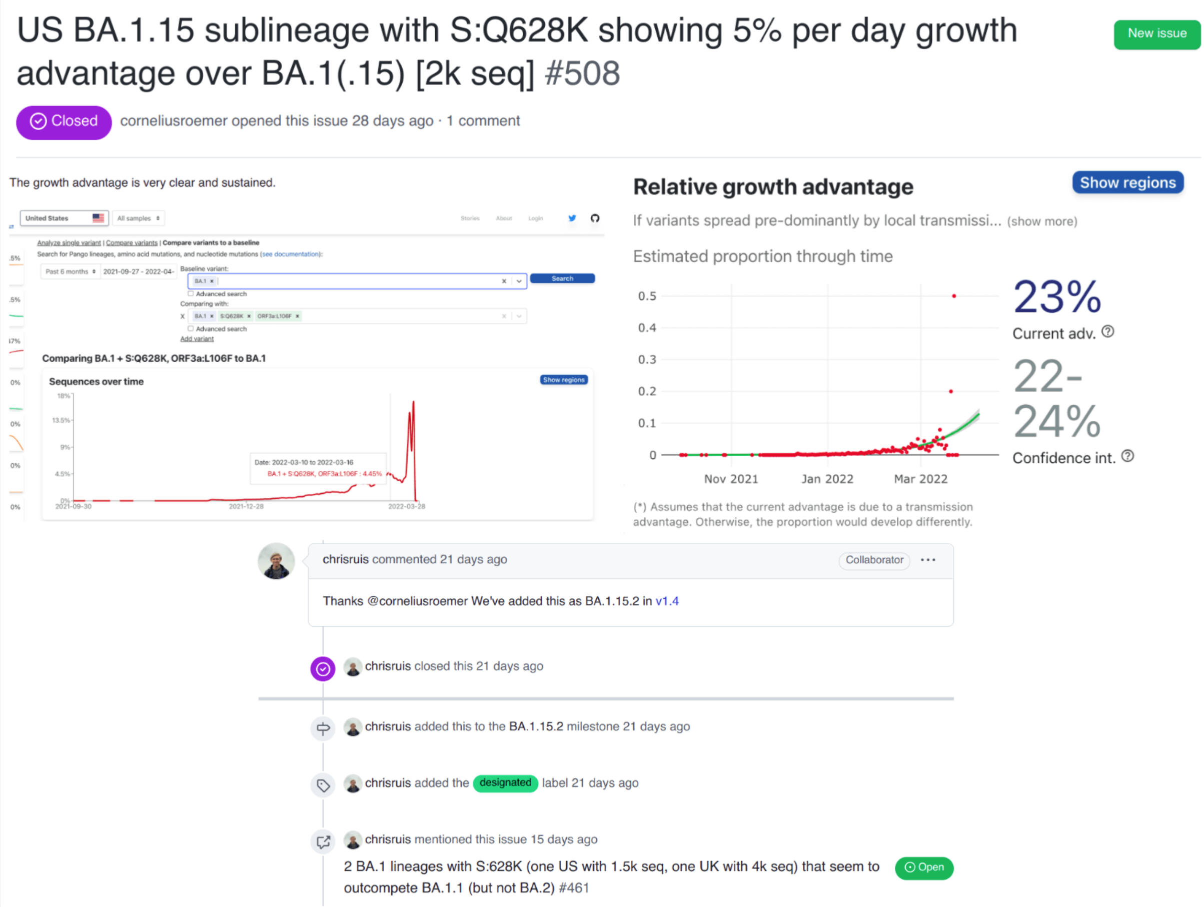
(c) Diffusion Odd Ratio.

Figure 3.21: Results of lineage specific analysis on BA.2.12 for the period 04/03/2022 - 31/03/2022 in USA. Data from Nextstrain dated 19/04/2022.

From the results summarized in Figure 3.21, a substantial increase in the prevalence of `L452Q` within the BA.2.12 lineage can be clearly observed. Indeed, as can be seen from both the Diffusion Odd Ratio graph and the Diffusion Heatmap, the amino acid change under analysis appears to be associated with a clear growth advantage.

Omicron BA.1.15.2 lineage

This use case considers a proposal for a potential **sub-lineage of BA.1.15** published in early April 2022.



US BA.1.15 sublineage with S:Q628K showing 5% per day growth advantage over BA.1(.15) [2k seq] #508

Closed corneliusroemer opened this issue 28 days ago · 1 comment

The growth advantage is very clear and sustained.

New issue
United States
All samples
Stories
About
Login
Analyze single variant | Compare variants | Compare variants to a baseline
Search for Pango lineages, amino acid mutations, and nucleotide mutations (see documentation):
Past 6 months
2021-09-27 - 2022-04-
Baseline variant:
BA.1
Search
Advanced search
Comparing with:
BA.1
S:Q628K
ORF3a:L106F
Advanced search
Add variant
Comparing BA.1 + S:Q628K, ORF3a:L106F to BA.1
Sequences over time
Show regions
18%
13.5%
9%
4.5%
0%
2021-09-30
2021-12-28
2022-03-28
Date: 2022-03-10 to 2022-03-16
BA.1 + S:Q628K, ORF3a:L106F : 4.45%
Relative growth advantage
Show regions
If variants spread pre-dominantly by local transmissi... (show more)
Estimated proportion through time
0.5
0.4
0.3
0.2
0.1
0
Nov 2021
Jan 2022
Mar 2022
23%
Current adv.
22-24%
Confidence int.
(*) Assumes that the current advantage is due to a transmission advantage. Otherwise, the proportion would develop differently.

chrisruis commented 21 days ago Collaborator ···

Thanks @corneliusroemer We've added this as BA.1.15.2 in v1.4

chrisruis closed this 21 days ago

chrisruis added this to the BA.1.15.2 milestone 21 days ago

chrisruis added the designated label 21 days ago

chrisruis mentioned this issue 15 days ago

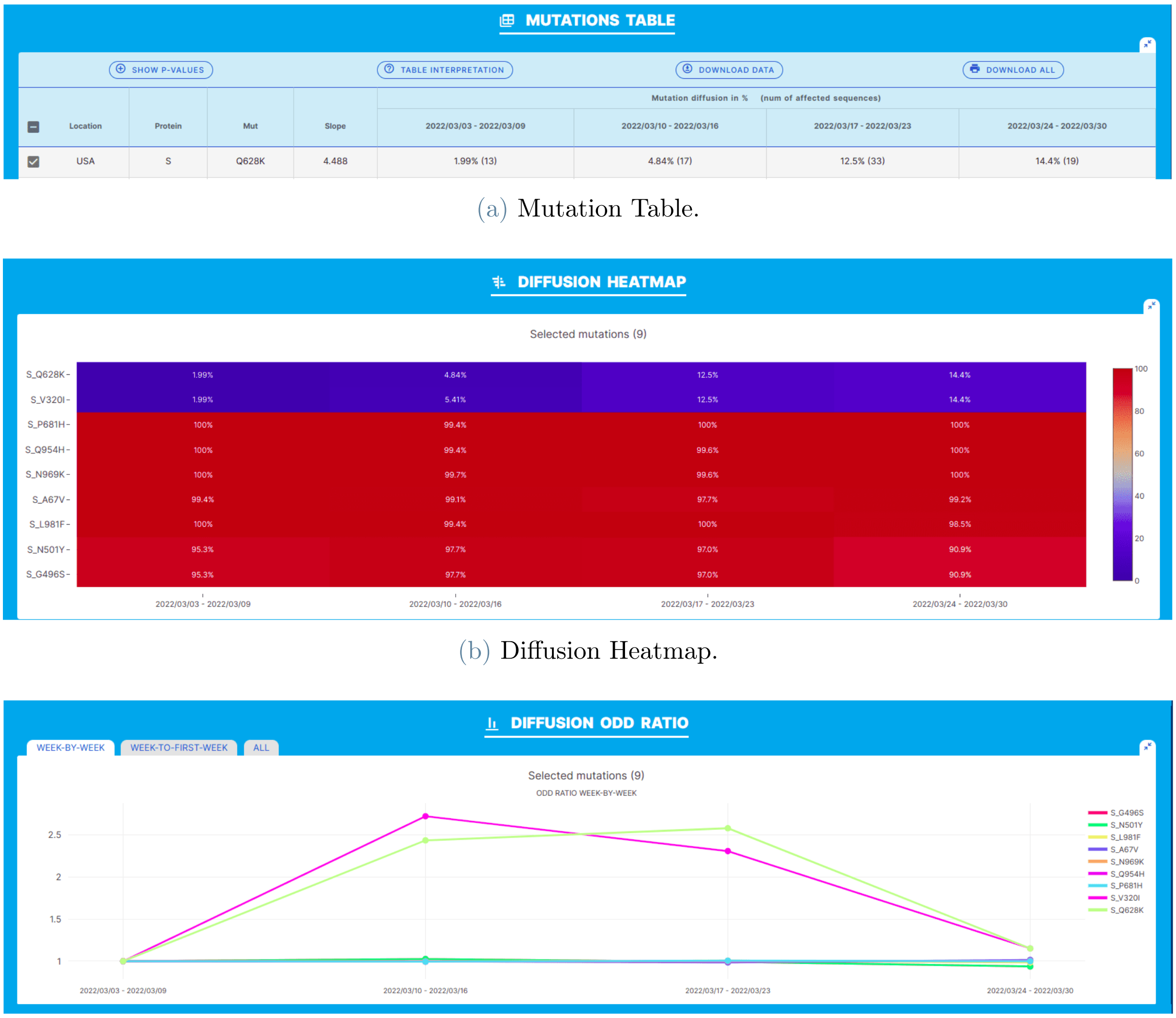
2 BA.1 lineages with S:628K (one US with 1.5k seq, one UK with 4k seq) that seem to outcompete BA.1.1 (but not BA.2) #461 Open

Figure 3.22: Issue 508 from Pango designation [8] which led to the definition of the BA.1.15.2 sub-lineage of SARS-Cov-2.

As reported in Figure 3.20, the novel lineage is defined by the amino acid change `Q628K` in the spike glycoprotein, according to the submitter. Also in this case, the proposal then resulted in the definition of a new sub-lineage, namely BA.1.15.2.

Also in this case Variant Hunter makes it easy to validate the issue. Indeed, it is sufficient to consider a *Lineage Specific* analysis on the parent lineage (i.e., BA.1.15) in the geographic location (USA) and interval of time (March 2022) where the novelty first emerged.

Again, the Nextstrain data set available as of 19/04/2022 was used, which at the time still did not consider the existence of BA.1.15.2.



MUTATIONS TABLE

SHOW P-VALUES | TABLE INTERPRETATION | DOWNLOAD DATA | DOWNLOAD ALL

⊟	Location	Protein	Mut	Slope	Mutation diffusion in % (num of affected sequences) 2022/03/03 - 2022/03/09	 2022/03/10 - 2022/03/16	 2022/03/17 - 2022/03/23	 2022/03/24 - 2022/03/30
☑	USA	S	Q628K	4.488	1.99% (13)	4.84% (17)	12.5% (33)	14.4% (19)

(a) Mutation Table.

DIFFUSION HEATMAP

Selected mutations (9)

	2022/03/03 - 2022/03/09	2022/03/10 - 2022/03/16	2022/03/17 - 2022/03/23	2022/03/24 - 2022/03/30
S_Q628K	1.99%	4.84%	12.5%	14.4%
S_V320I	1.99%	5.41%	12.5%	14.4%
S_P681H	100%	99.4%	100%	100%
S_Q954H	100%	99.4%	99.6%	100%
S_N969K	100%	99.7%	99.6%	100%
S_A67V	99.4%	99.1%	97.7%	99.2%
S_L981F	100%	99.4%	100%	98.5%
S_N501Y	95.3%	97.7%	97.0%	90.9%
S_G496S	95.3%	97.7%	97.0%	90.9%

(b) Diffusion Heatmap.

DIFFUSION ODD RATIO

(c) Diffusion Odd Ratio.

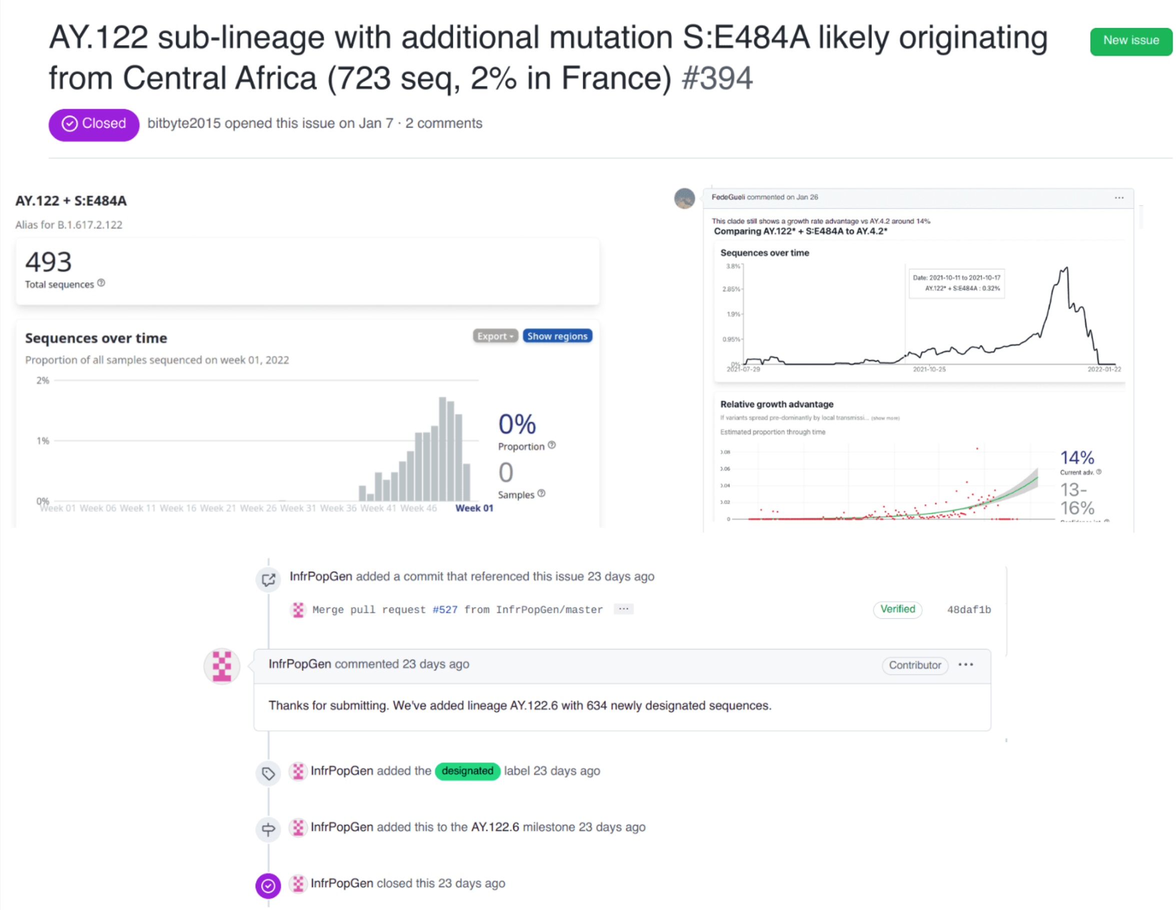
Figure 3.23: Results of lineage specific analysis on BA.1.15 for the period 03/03/2022 - 30/03/2022 in USA. Data from Nextstrain dated 19/04/2022.

From the results summarized in Figure 3.23, a substantial increase in the prevalence of L452Q within the BA.1.15 lineage can be observed both from the Mutation Table and the Diffusion Heatmap. Interestingly, we also observe a similar increase in the prevalence of S:V320I, an amino acid change that, at the time of writing, is not associated with any sub-lineage of BA.1.15.

Both mentioned mutations show a similar relative growth advantage over the amino acid changes that define the parent lineage. As a matter of fact, from Figure 3.23c, it can be seen that after a rapid surge in frequency (from 1% to 12%), their diffusion seems to stabilize around 15% percent, in the last week under analysis (i.e., their odd ratio is almost unitary).

Delta AY.122.6 lineage

Based on GISAID data as of 06/02/2022, this use case considers a proposal for a potential **sub-lineage of AY.122.6** published on January 2022.



AY.122 sub-lineage with additional mutation S:E484A likely originating from Central Africa (723 seq, 2% in France) #394

Closed bitbyte2015 opened this issue on Jan 7 · 2 comments

FedeGueli commented on Jan 26

This clade still shows a growth rate advantage vs AY.4.2 around 14%

Comparing AY.122* + S:E484A to AY.4.2*

InfrPopGen added a commit that referenced this issue 23 days ago

```
Merge pull request #527 from InfrPopGen/master
```

Verified 48daf1b

InfrPopGen commented 23 days ago

Thanks for submitting. We've added lineage AY.122.6 with 634 newly designated sequences.

InfrPopGen added the designated label 23 days ago

InfrPopGen added this to the AY.122.6 milestone 23 days ago

InfrPopGen closed this 23 days ago

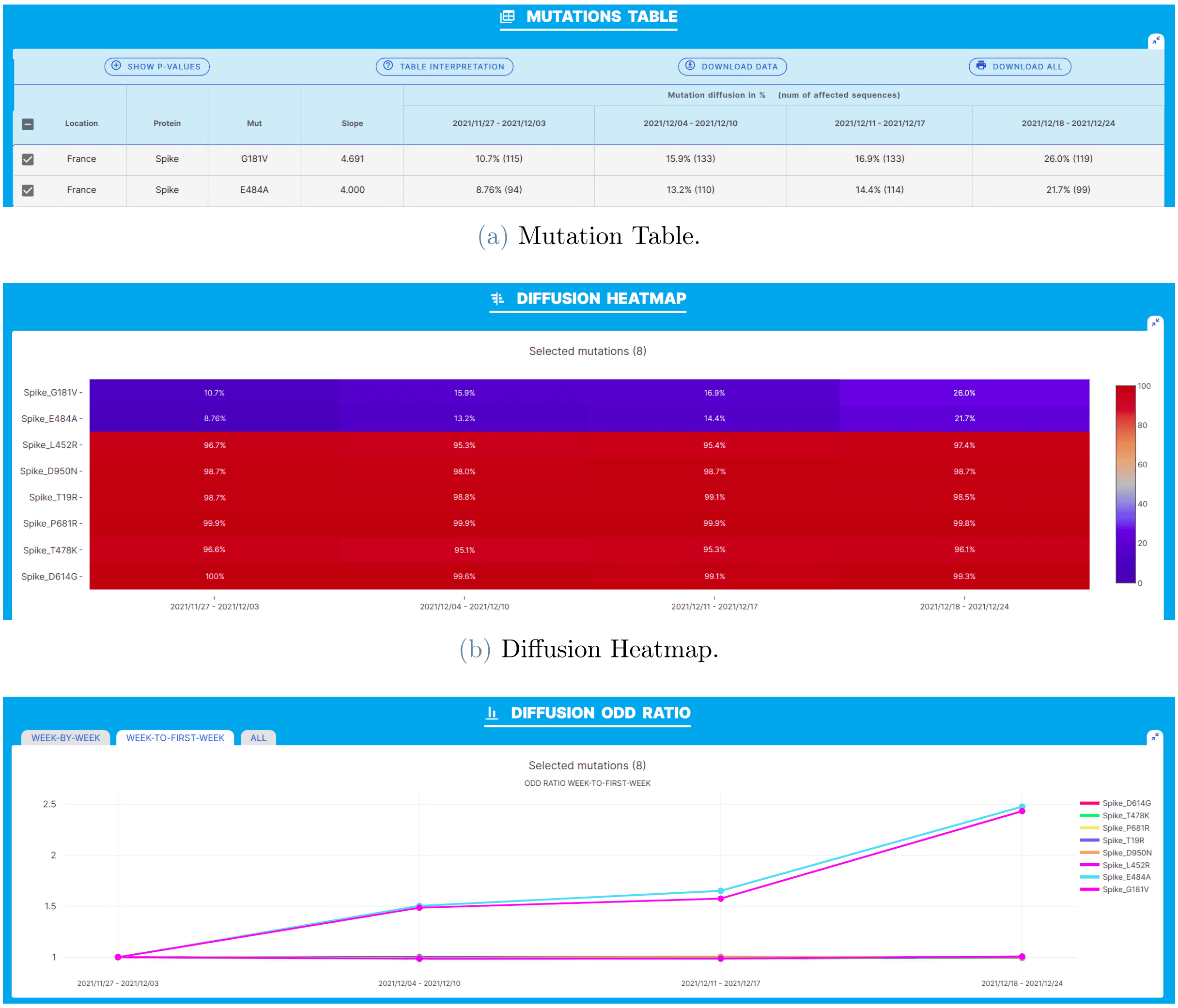
New issue
AY.122 + S:E484A
Alias for B.1.617.2.122
493
Total sequences
Sequences over time
Export
Show regions
Proportion of all samples sequenced on week 01, 2022
2%
1%
0%
Week 01 Week 06 Week 11 Week 16 Week 21 Week 26 Week 31 Week 36 Week 41 Week 46
Week 01
0%
Proportion
0
Samples
Sequences over time
3.8%
2.85%
1.9%
0.95%
0%
2021-07-29
2021-10-25
2022-01-22
Date: 2021-10-11 to 2021-10-17
AY.122* + S:E484A : 0.32%
Relative growth advantage
If variants spread pre-dominantly by local transmissi... (show more)
Estimated proportion through time
0.08
0.06
0.04
0.02
0
14%
Current adv.
13-16%
Contributor

Figure 3.24: Issue 394 from Pango designation [8] which led to the definition of the AY.122.6 sub-lineage of SARS-Cov-2.

As summarized in Figure 3.20, the novel lineage is defined by the amino acid change `E484A` in the spike glycoprotein, according to the submitter. This proposal resulted in the definition of the new sub-lineage AY.122.6.

The data considered in this example are prior to the reclassification with the new Pangolin schema updated with the addition of the new sub-lineage.

Verifying the report with Variant Hunter is again very simple. To look for confirmation, we can consider the *Lineage Specific* analysis on the AY.122 lineage in the geographic location (France) and interval of time (November-December 2022) where the novelty first emerged.



MUTATIONS TABLE

					Mutation diffusion in % (num of affected sequences)			
☐	Location	Protein	Mut	Slope	2021/11/27 - 2021/12/03	2021/12/04 - 2021/12/10	2021/12/11 - 2021/12/17	2021/12/18 - 2021/12/24
☑	France	Spike	G181V	4.691	10.7% (115)	15.9% (133)	16.9% (133)	26.0% (119)
☑	France	Spike	E484A	4.000	8.76% (94)	13.2% (110)	14.4% (114)	21.7% (99)

(a) Mutation Table.

(b) Diffusion Heatmap.

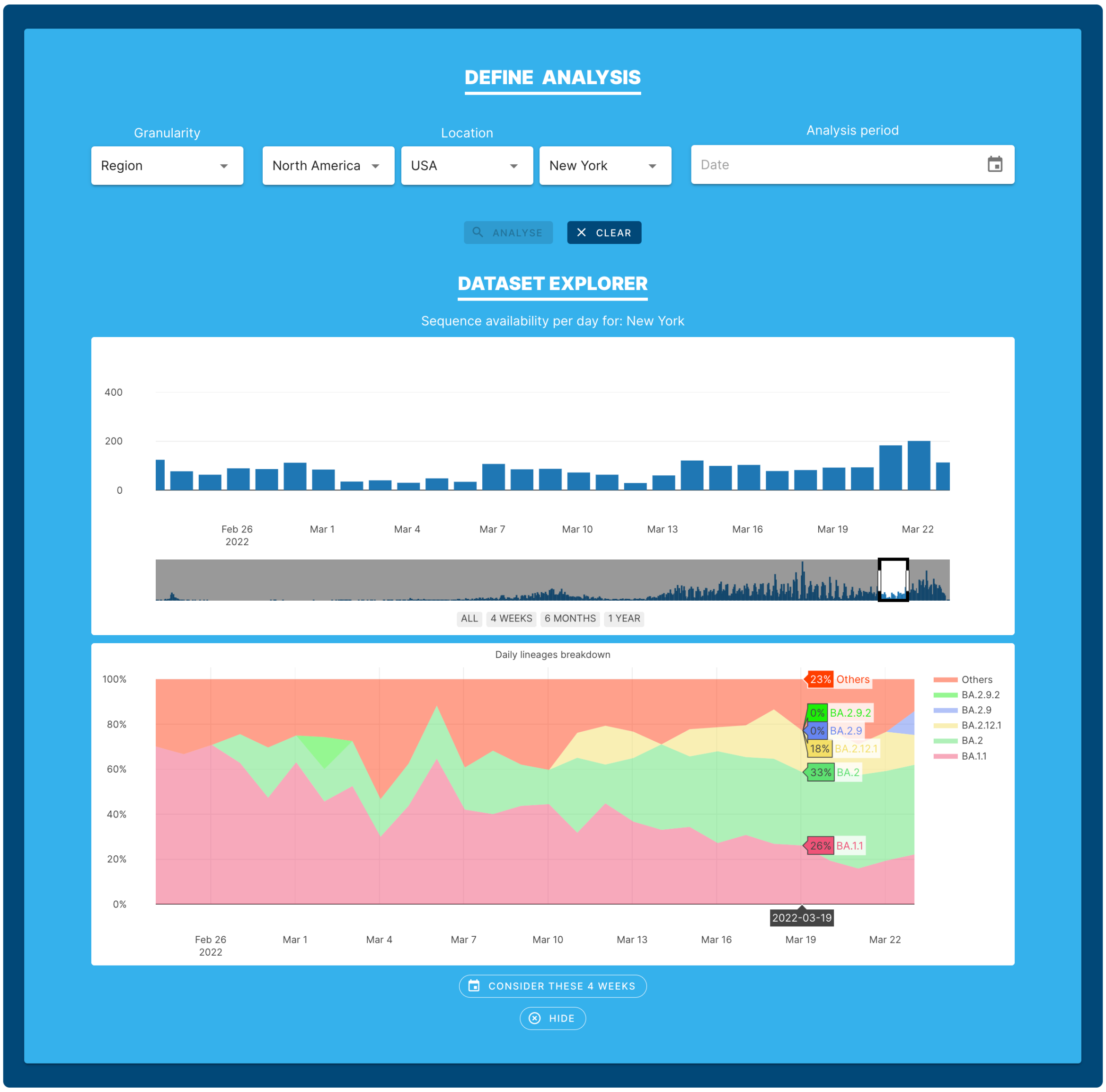
(c) Diffusion Odd Ratio.

Figure 3.25: Results of lineage specific analysis on AY.122 for the period 27/11/2021 - 24/12/2021 in France. Data from GISAID dated 06/02/2022.

From Figure 3.25 we can immediately notice a substantial increase in the prevalence of two amino acid changes: `S:E484A` and `S:181V`. According to the Diffusion Heatmap in Figure 3.25b both changes show a consistent increase in prevalence (from 10% to 26%) in the of time range included in the analysis.

3.3. Sequencing Dataset Explorer

The *Dataset Explorer* feature allows you to examine the sequence dataset directly from the analysis definition area. This functionality is available both for the *Lineage Independent* and the *Lineage Specific* tab. Its purpose is to support the choice of the time frame to be considered for analysis.



DEFINE ANALYSIS

Granularity: Region

Location: North America, USA, New York

Analysis period: Date

DATASET EXPLORER

Sequence availability per day for: New York

Figure 3.26: Dataset Explorer for the Lineage Independent tab showing sequence counts and lineage breakdown in New York. Data from GISAID dated 04/06/2022.

The Dataset Explorer is particularly useful to avoid selecting periods in which sequences are present in small numbers and thus provide poorly supported patterns. Indeed, given

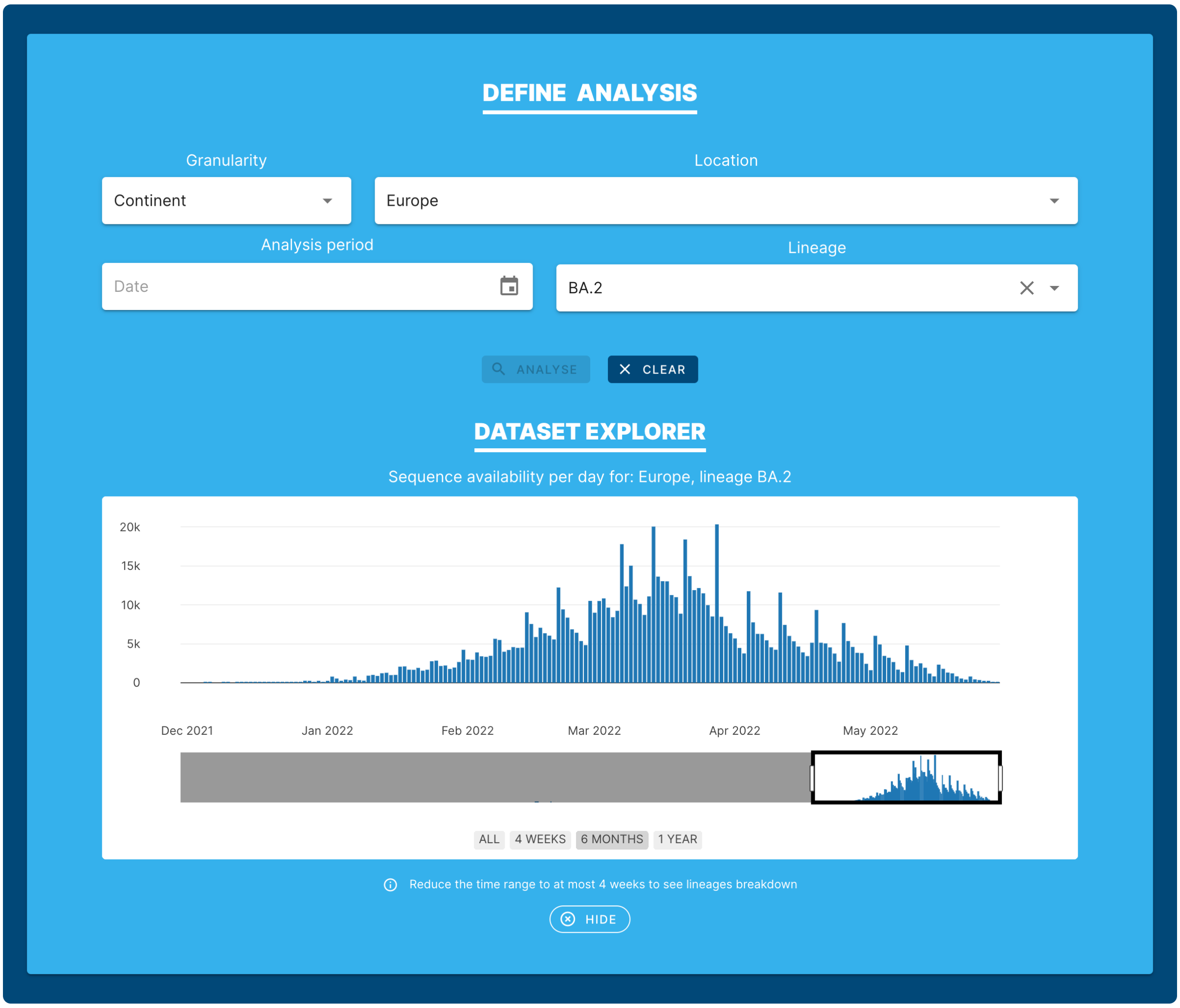
a location and possibly also a lineage, it allows to display a bar chart and a stacked area chart. Specifically, the former shows for each day the number of sequences in the dataset that have the specified characteristics.
As visible from Figure 3.26, a toolbar at the bottom allows the user to move through time or select a window of specific size (4 weeks, 6 months, 1 year).

When used in four-weeks mode, the explorer also shows a plot of the distribution of daily lineages. This allows one to:

- discover which lineage is prevalent in a given time period at a given location;
- analyze the trend of one lineage with respect to one or more other lineages. Indeed the charts are fully flexible and allow to select only specific lineages of interest.

For example, in the case in Figure 3.26 in *New York* on *March 19th, 2022*, 33% of the sequences were *BA.2*, 26% *BA.1.1*, 18% *BA.2.12.1* and 23% other minor lineages (i.e., lineages involving less than 10% of the sequences collected on that day).

Instead, in Figure 3.27, the Dataset Explorer is shown in action in case a lineage is also selected. Specifically, the example depicts the number of sequences associated with lineage BA.2 in Europe in a 6-month time window ending on 29/05/2022.



DEFINE ANALYSIS

Granularity: Continent

Location: Europe

Analysis period: Date

Lineage: BA.2

DATASET EXPLORER

Sequence availability per day for: Europe, lineage BA.2

Reduce the time range to at most 4 weeks to see lineages breakdown

Figure 3.27: Dataset Explorer for the Lineage Specific tab showing sequence counts in Europe for BA.2. Data from GISAID dated 04/06/2022.

4 | Design and Implementation

Variant Hunter is implemented in the form of a web application that runs in standard browsers. Specifically, the application architecture includes three fundamental building blocks that communicate with each other. First, SARS-CoV-2 sequencing data are stored in aggregated form in a text-based database. Particular attention was devoted to the design of the schema in order to minimize the processing time and to optimize disk occupancy. Next, the backend part fetches sequence information from the data store and uses it to feed powerful analysis algorithms. The results of these analyses, including the frequencies of amino acid mutations of SARS-CoV-2, are made available through a set of APIs. Finally, the frontend accesses the exposed endpoints and, after further client-side processing, displays charts and statistics to the end user.

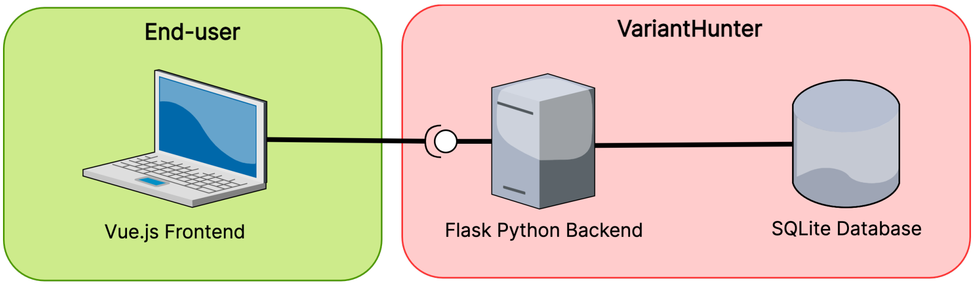


Figure 4.1: Overall logical design.

As the ultimate goal of Variant Hunter is to facilitate data analysis, the application was entirely designed with aspects such as flexibility, simplicity and user-friendliness in mind. The result is a *fast* application, capable of handling massive amounts of data and returning, in fractions of seconds, visualizations capable of conveying relevant information about pandemic trends.

This chapter details the design and implementation choices made in the software development process. Specifically, the first part presents the two largest collections of SARS-CoV-2 sequences available today (**which data are available**): GISAID [26] and GenBank [24]. More specifically, their structure, the preparation and the cleaning procedures, which allows to overcome possible data quality issues, are analyzed. Particular attention is also devoted to the introduction of the challenges that such a large amount of data entails, both in spatial and temporal terms.

Next, comes the presentation of the solution adopted for storing this information in aggregated form (**how data are stored**). It is then discussed the identification of the data analysis methods that have been adopted in the software in order to identify the mutations whose prevalence is significantly increasing in a considered analysis period (**how data are processed**).

Finally, to conclude the chapter, it is presented the graphical user interface of the tool, properly designed to address the specific needs of target users (**how data are presented**).

4.1. Available Sequencing Data

The landscape of initiatives dedicated to the collection and publication of viral sequences is a broad one.

Many institutions have joined the fight against COVID-19 and have begun to gather data on SARS-CoV-2. Some of them, such as NCBI's GenBank [24], pre-existed the COVID-19 pandemic and harbored thousands of viral species that pose a threat to humanity, including but not limited to Ebola and SARS [2]. On the other hand, other organizations have produced new data collections specifically dedicated to host SARS-CoV-2 data, such as GISAID [26]. Originally designed to collect influenza viral sequences, the latter soon became the predominant source for SARS-CoV-2 sequences.

At the time of writing, the two main collections of SARS-CoV-2 data are GISAID [26] and GenBank [24], which contain 11.2 and 4.8 million entries, respectively. These sequences have been obtained mainly through submissions from individual laboratories and batch submissions from large-scale sequencing projects, and their numbers are rapidly increasing as new sequences are continuously being deposited.

A first difference between the two considered repositories lies in the restrictions placed on their access. As a matter of fact, in the case of GISAID, the data are not publicly available: in order to access them, one is required to own an account, issued upon request after the acceptance of an agreement. Thanks to this controlled policy, GISAID has been able to attract more scientists, including those who reluctantly share data within completely open-source repositories [2]. On the other side, Genbank is a public data provider making its resources freely available through dedicated web interfaces.

Both of the considered providers deliver the data of interest for the development of this project in the form of metadata files, in .tsv format. Actually, these files describe the collected SARS-CoV-2 sequences by providing a rich set of information, including date and place of sampling, lineage of belonging, and the list of substitutions in amino acids.

Although they provide the same information, the structure adopted by the two organizations differs significantly in some aspects. The following sections presents the major differences between them and discuss the procedures required to solve data quality issues of the individual data sources. Regardless, the tool is able to deal with both formats and possibly also with data from other sources (e.g., private data sets) based on the same structure. More precisely, for what concerns GenBank data, the metadata processed according to the ncov [18] workflow by Nextstrain [13] are considered.

4.1.1. GISAID metadata

GISAID [26] represents the largest source of sequencing data available today for SARS-CoV-2. The tool user will have to autonomously retrieve the file, from the dedicated portal provided by GISAID, and feed it as input to the software at first startup.

The metadata file, provided in tsv format, was carefully examined using Python scripts. The goal of this analysis is to identify which data are available and determine the transformations to be performed to standardize the formats. Specifically, Table 4.1 details the main fields of interest to the application along with some examples.

The main data quality problems detected during the analysis are related to the following aspects:

Issue 1 Some rows have low sequence completeness (high percentage of unknown bases).

Issue 2 Collection dates have accuracy issues: they often miss the month or the day of collection.

Issue 3 Also geographical location data, including continent, country and region, suffer from accuracy problems. Sometimes the same region is encoded differently within the same file (e.g., Emily-Romagna vs Emilia-Romagna). Moreover, in few cases, names of cities are used instead of those of the regions (e.g., Castano Primo is not an italian region). In some other cases, location information is also incomplete: one or more among region, country or continent is missing.

Column	Name	Description
3	*Collection date*	Date in which the infected biological sample was collected. Takes the format `YYYY` or `YYYY-MM` or `YYYY-MM-DD`. *Example:* 2021-08-17
4	*Location*	Sample location in the format `CONTINENT`, `CONTINENT/COUNTRY` or `CONTINENT/COUNTRY/REGION` *Example:* Europe/Turkey
6	*Sequence length*	Number of nucleotides of the sequence. *Example:* 29772
11	*Pango lineage*	Sequence lineage description. *Example:* B.1.617.2
14	*AA Substitutions*	List of amino acid substitutions. *Example:* (N_D377Y,Spike_D950N,. . .)
20	*N-Content*	Percentage of unknown bases. *Example:* 0.000134354426978

Table 4.1: Description of fields of metadata.tsv from GISAID

The following expedient are taken to standardize information and exclude non-significant data:

Solution 1 The completeness of the sequences must be carefully analyzed for each row. This can be done by exploiting the *"Sequence length"* and *"N-Content"* fields. Specifically, the following rule is adopted for selecting the data to be imported into the tool and overcome Issue 1:

```
(29000 < sequence_length < 30000) and (n_content < 0.05)
```

Solution 2 Since the accuracy of the sampling date is considered essential, and given the impossibility of placing a sample on the time axis in the absence of day or month information (Issue 2), data with problems in this regard are automatically discarded.

Solution 3 Given that the analysis can be done at different levels of granularity (continental, national or regional analysis), the absence of precise location information (Issue 3) does not represent major problem. Incomplete data are imported into the tool anyway. For example, data that do not provide region information will be considered only for national and continental analyses.

Finally, issues with the naming of regions are simply ignored as they represent sporadic errors, often corrected in subsequent releases of the metadata file.

4.1.2. Nextstrain metadata

The metadata published by Nextstrain [13] represent the largest collection related to COVID-19 that is publicly available and freely usable. As previously explained, it is constituted by GenBank [24] data processed according to the ncov [18] workflow.

Also in this case, the metadata file is provided in tsv format, and has been examined through Python scripts.

In Table 4.2 the relevant fields for the development of Variant Hunter are explained along with some value examples.

The major data quality problems detected during the analysis are related to the following aspects:

Issue 1 Some rows have low sequence completeness (high number of unknown bases).

Issue 2 Sample collection dates have problems with completeness: they are often missing.

Issue 3 In this data source, geographic information is found to be more accurate from the point of view of consistency in the nomenclature employed. However, special attention should be paid to the *"Division"* field, which often coincides with *"Country"* when its value is not known. In some other cases, location information is also incomplete: one or more among region, country or division is missing.

Issue 4 The character *'?'* and the strings *'None'* and *'Unassigned'* are interchangeably used to represent the fact that the value for *"Pango lineage"* is not known. This represents an accuracy problem.

Column	Name	Description
5	*Date*	Date in which the infected biological sample was collected. Takes the format `YYYY-MM-DD`. *Example:* 2021-03-15
6	*Region*	Name of the continent where the sample was collected. *Example:* Europe
6	*Country*	Name of the country where the sample was collected. *Example:* Italy
6	*Division*	Geographic area of finer granularity. *Example:* Lombardy
14	*Length*	Number of nucleotides of the sequence. *Example:* 29912
19	*Pango lineage*	Sequence lineage description. *Example:* B.1.1.7
30	*Missing data*	Number of unknown bases. *Example:* 19
48	*AA Substitutions*	List of amino acid substitutions. *Example:* (N:D377Y,S:D950N,. . .)

Table 4.2: Description of fields of metadata.tsv from Nextstrain

In order to normalize the source and exclude non-significant data, the following steps were taken:

Solution 1 Also for Nextstrain, the completeness of the sequences must be analyzed for each row. This can be done by exploiting the *"Missing data"* and *"Length"* fields. Specifically, the following rule is adopted to select the data to be imported into the tool and overcome Issue 1:

```
(29000 < length < 30000) and (missing_data/length < 0.05)
```

Solution 2 Due to the importance of sampling date accuracy and given the impossibility of placing a sample on the time axis in its absence (Issue 2), data with date-related errors are ignored.

Solution 3 Given that the analysis can be done at different levels of granularity (continental, national or regional analysis), the absence of precise location information (Issue 3) does not represent major problem. For example, data that do not provide information related to the Division will be considered only for national and continental analyses.

Moreover, the *"Division"* field is considered as not given if it does not provide additional knowledge than *"Country"* and *"Region"*.

Solution 4 To overcome Issue 4, the labels are re-mapped as follows:

```
('?','Unknown','None') => 'None'
```

4.1.3. Challenges

As previously mentioned, GISAID [26] and GenBank [24], provide metadata related to 11.2 and 4.8 million of sequences, respectively. Furthermore, it should be considered that the numbers are growing very rapidly as new data are being deposited daily by hundreds of laboratories around the world.

This clearly poses challenges in their management. Such large volumes of data require special care not only in how they are stored, but also in the way they are processed by the analysis algorithms.

In addition, as discussed in Section 4.1.1 and Section 4.1.2, the two files have significant differences in data structure and semantics. Solutions to solve the data quality issues of individual data sources have already been discussed previously. Still, it is necessary to identify a standardized **target schema** that allows the two data resources to be used in an *interchangeable fashion*. Therefore, schema integration procedures are of paramount importance.

In summary, key design requirements are:

Requirement 1 Identification of a global integrated scheme, together with the definition of the procedures to perform conflicts resolution;

Requirement 2 Design and implementation of data structures capable of being efficient in terms of access time and storage space required;

Requirement 3 Design and implementation of efficient processing algorithms.

4.2. Database

Starting from the requirements identified in Section 4.1.3, in this chapter the database design and integration procedures are presented. More specifically, the first part discusses the aspects of schema reconciliation and conflict resolution. After that, the overall schema and the procedures for data extraction, transformation and loading (ETL) are presented. Finally, the SQL database engine used for the implementation is introduced.

4.2.1. Schema alignment

The first step toward the definition of the final database that feeds Variant Hunter is the schema integration. Indeed, the two sources contain different numbers of attributes; they use different names to refer to attributes with the same semantic (e.g., *"Sequence Length"* in GISAID vs. *"Length"* in Nextstrain) and they apply different semantics for attributes with the same name (e.g., *"Region"* in GISAID represents the lower lever of granularity for the location, while in Nextstrain is used to indicate the continent).

The *traditional approach* for schema alignment consists of three steps: creating a mediated schema, attribute matching, and schema mapping, depicted in Figure 4.2. [32]

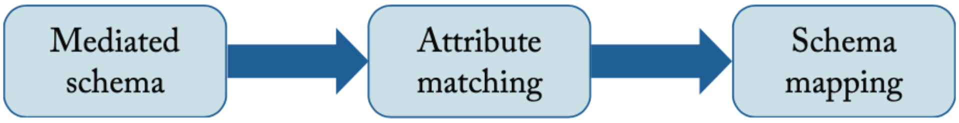


Figure 4.2: Traditional schema alignment: three steps. [32]

Mediated schema

First, a mediated schema is created to provide a unified virtual view of the disparate sources and capture the salient aspects of the domain under consideration. [32]

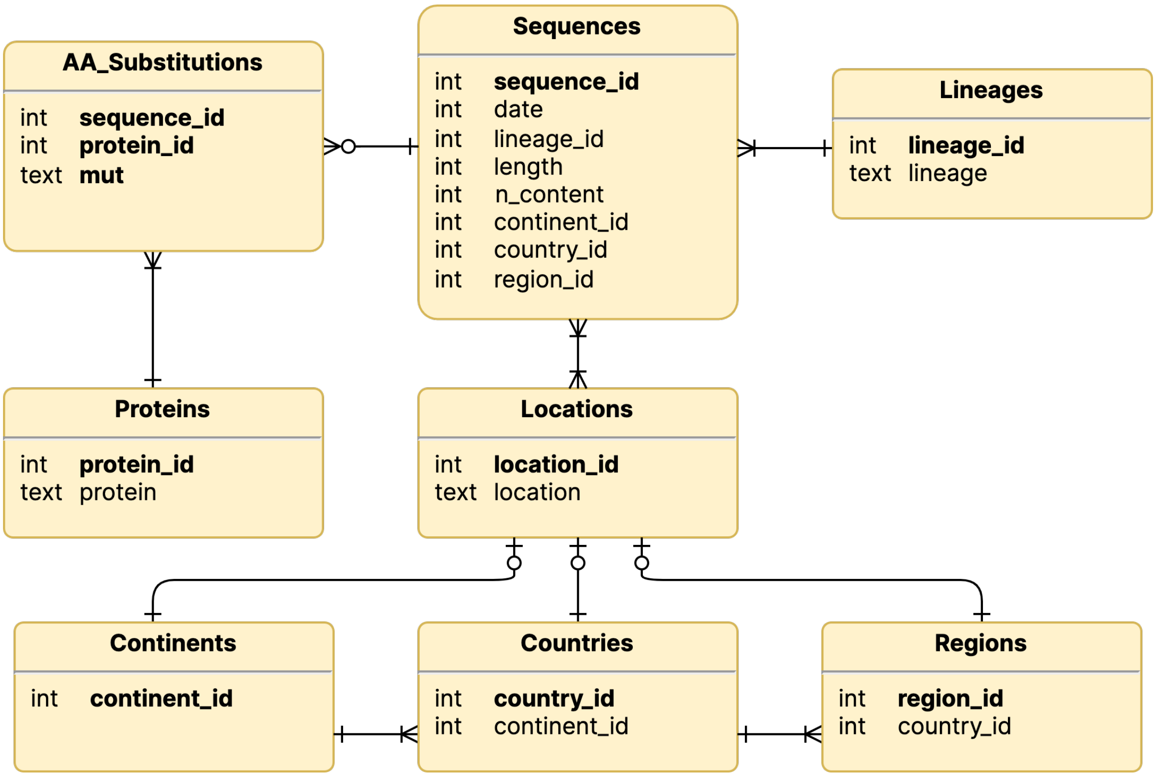


Figure 4.3: Mediated schema for the data sources

As shown in Figure 4.3, the schema contains the following tables:

- **Sequences** for sequence-related information such as sampling date, length and missing basis;
- **AA_Substitutions** for information about amino acid substitutions detected in sequencing;
- **Proteins** and **Lineages** for information about proteins and lineages, respectively;
- **Locations**, **Continent**, **Country**, and **Region** for information related to the geographical areas.

As typical for a unified view, the mediated schema contains knowledge that is not directly present in both the sources. For example, it contains information about the percentage of unknown bases (*"N_content"* attribute of Sequences table) that is not (directly) provided by Nextstrain. Note also that the global schema does not contain all the knowledge from each source. For example, GISAID provides host information and a very large number of sequencing-related attributes, but they are not included in the representation as they are not relevant for the system to be.

A final observation to be made is related to the fact that the mediated schema in Figure 4.3 is not the one actually used by the application. Indeed, the former is solely the starting point from which the latter can be constructed.

Attribute matching and schema mapping

Next, the attributes of each source schema are matched with the corresponding attributes of the mediated schema [32]. In many cases, attribute matching is one-to-one; however, sometimes an attribute in the mediated schema matches the combination of multiple attributes in the source schema and vice versa. For example, the combination of the attributes *"Length"* and *"Missing data"* in Nextstrain can be traced back to the *"N_content"* value in the mediated schema. Instead, in the case of GISAID, the value for *"N_content"* is directly provided.

The conflicts that need to be addressed in the context of related concept identification are the following:

Name conflicts: refers to situations in which *different names* are used to represent the *same concept*. It is the most frequent conflict here.

For example, GISAID uses the terms *"Collection date"* and *"Sequence length"*, while Nextstrain employs *"Date"* and *"Length"* respectively to represent the same concepts.

Structure conflicts: refers to situations in which the *same concept* is represented using *different structures*.

In the present case, GISAID uses a single attribute for the representation of the location (*"Location"*), while Nextstrain splits it into three different fields (*"Region"*,*"Country"* and *"Division"*)

Included under this umbrella is also the different format adopted by the two sources to encode amino acid substitutions. In the GISAID case the format `<protein>_<mutation>` is used; Nextstrain, on the other hand, represents them as `<protein>:<mutation>`

Semantic conflicts: refers to situations in which *same names* are used to represent *different concepts*.

As previously observed, a glaring example present here relates to the *"Region"* field. GISAID uses it to refer to the lowest level of granularity (an area of a country), while Nextstrain employs this name to refer to the highest one (an area of the world, i.e., a continent)

Based on the attribute correspondences, a schema mapping between each source schema and the mediated one is then constructed. These mappings specify the semantic relationships between the contents of the different data sources and are used to reformulate a query on the mediated schema into specific queries on the underlying data sources [32].

The result of the conflict resolution and the mapping analysis is reported in Table 4.3.

Aligned Knowledge	**GISAID Knowledge**	**Nextstrain Knowledge**
Sample date	Directly available as *Collection date*	Directly available as *Date*
Sample continent	Extracted from *Location* considering the format `CONTINENT/COUNTRY/REGION`	Directly available as *Region*
Sample country	Extracted from *Location* considering the format `CONTINENT/COUNTRY/REGION`	Directly available as *Country*
Sample region	Extracted from *Location* considering the format `CONTINENT/COUNTRY/REGION`	Directly available as *Division*
Sample length	Directly available as *Sequence length*	Directly available as *Length*
Associated lineage	Directly available as *Pango lineage*	Directly available as *Pango lineage*
N-Content	Directly available as *N-Content*	Computed as *Missing data/Length*
Proteins of amino acid substitutions	Extracted from *AA Substitutions* considering the format `<protein>:<mutation>`	Extracted from *AA Substitutions* considering the format `<protein>_<mutation>`

Mutations of amino acid substitutions	Extracted from *AA Substitutions* considering the format `<protein>:<mutation>`	Extracted from *AA Substitutions* considering the format `<protein>_<mutation>`

Table 4.3: Knowledge matching

4.2.2. Database design

The schema resulting from the integration of the two data sources is only the starting point for the development of the application's final database.

The original, and still predominant, approach to data analysis is the definition and creation of a centralised database called a data warehouse [1]. In the model adopted in this project, all the data required to perform specific data analyses are copied into a single DBMS that pre-computes much of the information and intermediate results so as to make a typical analysis running in less than a second on a common computer.

As illustrated in Figure 4.4, the general idea is to process data in three subsequent steps.

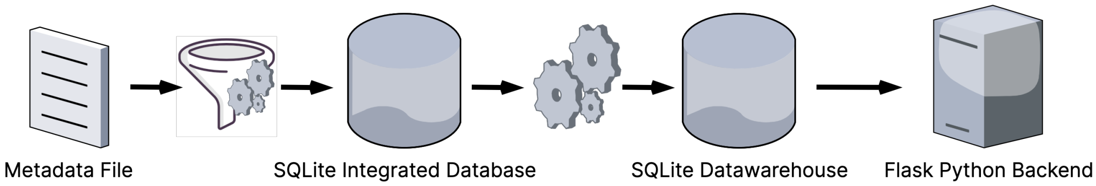


Figure 4.4: Overall data flow: three steps.

At first, the data are extracted from the considered source (i.e. the metadata.tsv file of GISAID [26] or Nextstrain [13]) and stored in an initial database structured according to the integrated schema previously identified in Section 4.2.1.

At this point, starting from the latter, we proceed with the aggregation of data and the computation of statistics. For example, it is at this stage that the number of sequences having a given mutation is computed for each continent, for each country and for each region.

These results are then stored in a new database, specifically designed to accommodate the requirements identified in Section 4.1.3.

Finally, the backend engine, upon receiving a request for analysis, proceeds by extracting the data of interest from the warehouse and processing them further. The result is the statistics that enables *variant hunting*.

In summary, the warehouse, as a coherent 'global snapshot' of sequencing data, is used to perform so-called *decision support* or *online analytical processing* (OLAP) queries, i.e. queries that examine aggregated features of the data to help take decisions [1], which in this context are related to public safety.

Data warehouse design process

First, it is necessary to define the relevant facts for the application, defined as concepts of interest for the analysis process. Based on the discussion in Chapter 3, it is clear that there are two concepts of interest in this domain:

Sequences: we are interested in analyzing the number of sequences collected in a given time period, for a given area and possibly in the context of a given lineage. Through this information then it will be possible to display the number of sequences collected and the lineage decomposition within the Dataset Explorer functionality. They will also be exploited for the purpose of computing diffusion statistics related to amino acid mutations.

AA Substitutions: we are interested in analyzing the number of times a given mutation is present in given time frame, for a given location, and possibly considering a specific lineage. By means of this information, it will be possible to compute the prevalence of mutations and derive other meaningful statistics.

Figure 4.5 and Figure 4.6 depict the attribute trees and the derived fact schemas for the two considered facts.

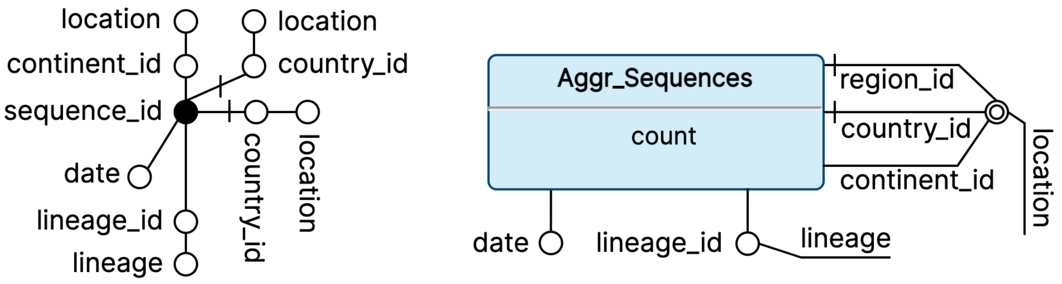


Figure 4.5: Attribute tree and derived fact schema for Sequences fact.

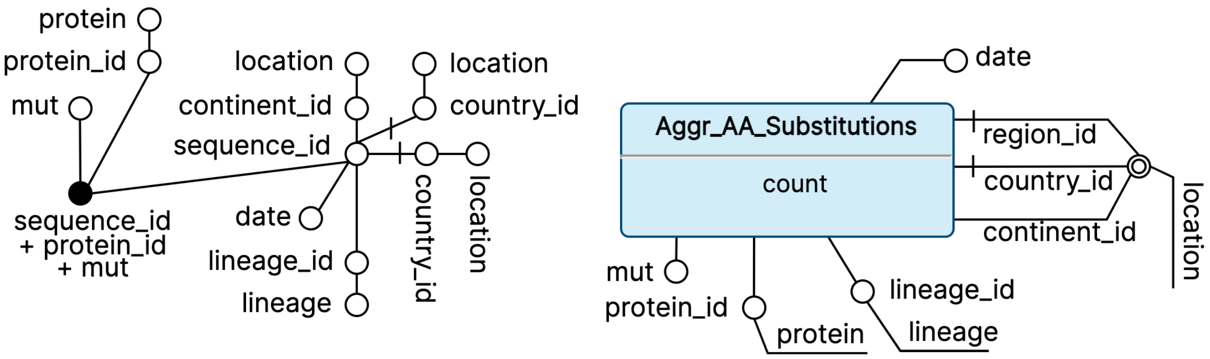


Figure 4.6: Attribute tree and derived fact schema for AA Substitutions fact.

Finally, the final logical schema of the application is shown in Figure 4.7. Specifically, it consists of the following tables:

- **Aggr_Sequences** for aggregated sequence-related information. It stores the number of collected SARS-CoV-2 sequences for every tuple `(Date,Lineage,Location)` where: `Date` represents the date of collection, `Lineage` represents the lineage associated with the sequences, and `Location` represents either a region, a country or a continent;
- **Aggr_AA_Substitutions** for aggregated information about amino acid substitutions detected in sequencing. It stores the number of collected SARS-CoV-2 sequences for every tuple `(Date,Lineage,Location,Protein,Mut)` where: `Date` represents the date of collection, `Lineage` represents the lineage associated with the sequences, `Location` represents either a region, a country or a continent, and the pair `Protein:Mut` represents the considered amino acid substitution;
- **Proteins** and **Lineages** for information about proteins and lineages, respectively;
- **Lineages_Characteristics** to store the characterising mutations for each lineage. These are identified as the mutations that are present in at least 50% of the lineage sequences;
- **Locations**, **Continent**, **Country**, and **Region** for information related to the geographical areas.

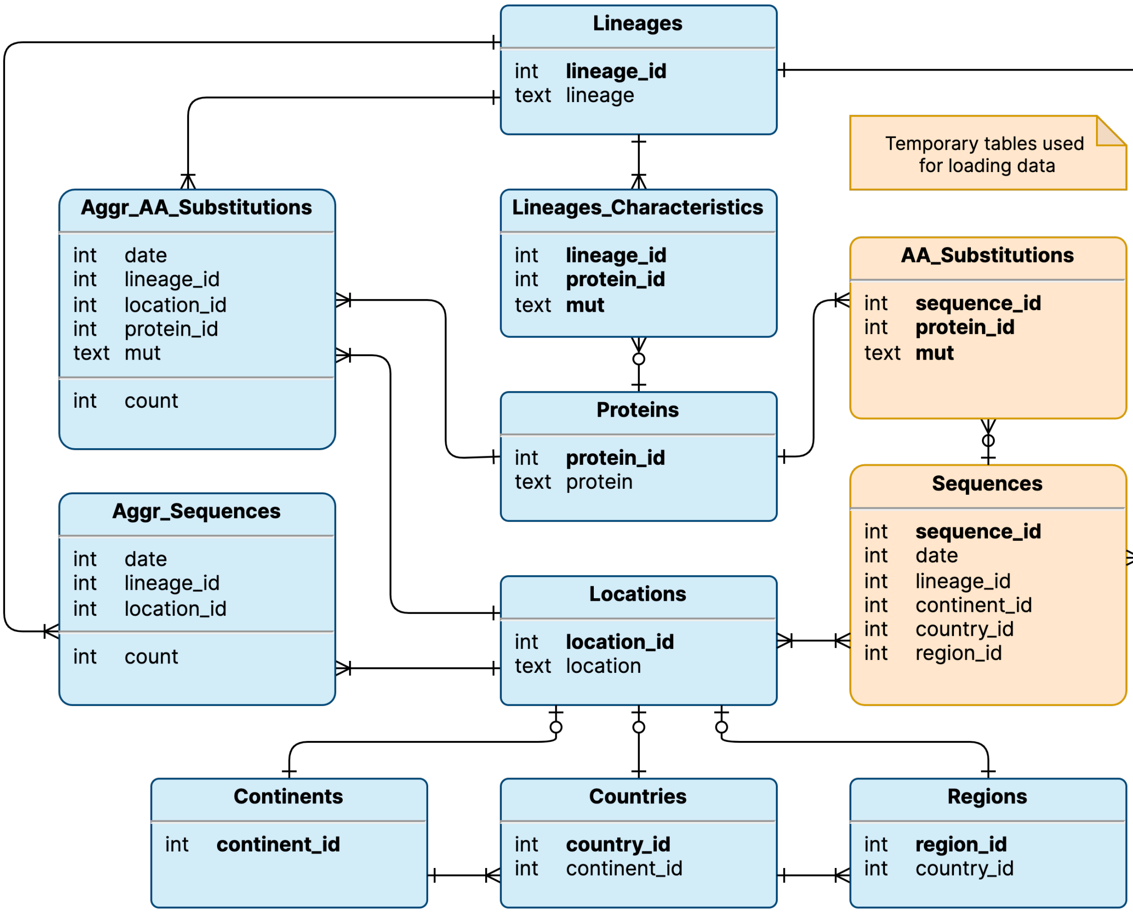


Figure 4.7: Final logical schema together with the temporary tables generated by parsing the metadata.tsv file. The latter are deleted once the aggregated data and the lineages characterizations have been computed.

Technical details

The present design was implemented using SQLite, an open-source «*C-language library that implements a small, fast, self-contained, high-reliability, full-featured, SQL database engine*» [7].

Unlike most other SQL databases, SQLite does not have a separate server process: it is an embedded database engine. Indeed, it reads and writes directly to standard disk files. A complete SQL database, with multiple tables, views and indexes is contained within a single disk file.

Moreover, SQLite is the most widely used database in the world, with a growing number of applications and high-profile projects based on it [7].

4.3. Backend and Data Analysis

The backend part of Variant Hunter is written in Python using the **Flask** framework [20]. In detail, Flask is known as a *micro-framework* as it is lightweight and aims to keep the core simple but extensible. Among its core functionalities, particularly useful in this project, are routing, request handling, and the template engine. Noteworthy is also the usage of **Flask-RESTPlus**, a Flask extension that adds support to quickly build REST APIs [22].

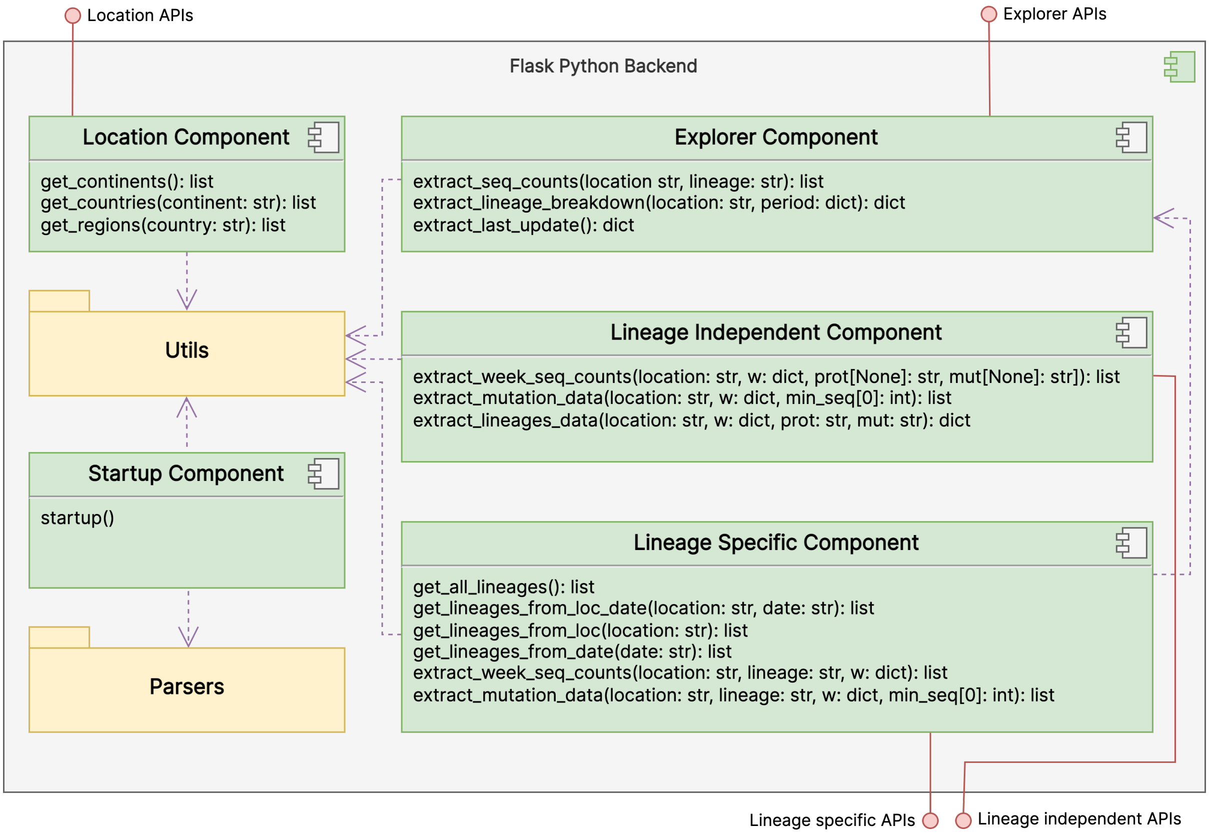

Location APIs
Explorer APIs
Flask Python Backend
Location Component
get_continents(): list
get_countries(continent: str): list
get_regions(country: str): list
Explorer Component
extract_seq_counts(location str, lineage: str): list
extract_lineage_breakdown(location: str, period: dict): dict
extract_last_update(): dict
Utils
Lineage Independent Component
extract_week_seq_counts(location: str, w: dict, prot[None]: str, mut[None]: str]): list
extract_mutation_data(location: str, w: dict, min_seq[0]: int): list
extract_lineages_data(location: str, w: dict, prot: str, mut: str): dict
Startup Component
startup()
Lineage Specific Component
get_all_lineages(): list
get_lineages_from_loc_date(location: str, date: str): list
get_lineages_from_loc(location: str): list
get_lineages_from_date(date: str): list
extract_week_seq_counts(location: str, lineage: str, w: dict): list
extract_mutation_data(location: str, lineage: str, w: dict, min_seq[0]: int): list
Parsers
Lineage specific APIs
Lineage independent APIs

Figure 4.8: Simplified high-level backend diagram

The backend has been structured mainly into four sub-components, each of which exposes a number of APIs. As also depicted in Figure 4.8, they are as follows:

- Explorer component: it processes the information necessary for the sequencing data exploration functionality.
- Lineage Specific component: it produces the statistics related to lineage specific analyses.

- Lineage Independent component: it produces the statistics related to lineage independent analyses.
- Location component: a support component for managing information related to geographic locations.
- Startup component: a support component for managing the initial startup, including parsing and aggregation steps.

In the following, the main components are analyzed closely, along with the analysis methods employed in order to generate the data returned by the endpoints.

4.3.1. Explorer component

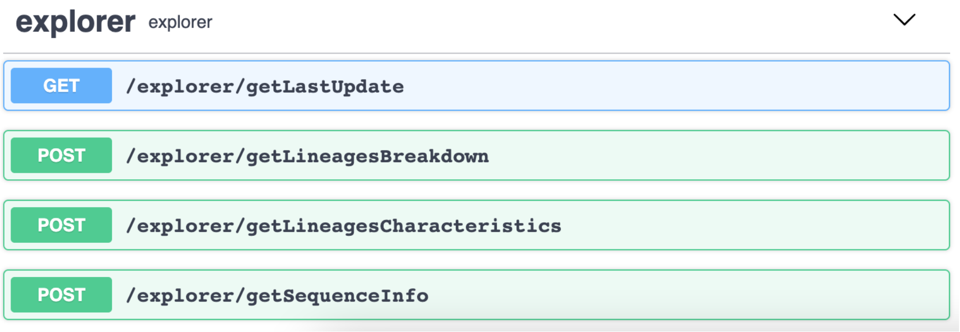
The Explorer component is primarily dedicated to the *Dataset Explorer* functionality, described in Chapter 3. As summarized by the Swagger documentation in Figure 4.9, it includes the following endpoints:

`/explorer/getSequenceInfo`: given a geographic area `A` (either a region, a country, or a continent) and possibly also a lineage `L`, it returns a list of pairs `(DAY, COUNT)` describing the number of sequences collected for each day, in `A` and characterized by `L` (if specified).

`/explorer/getLineagesBreakdown`: given a location `L` and a time period `[T1,T2]`, it returns a dictionary in which for every lineage detected in the period `[T1,T2]`, the number of daily affected sequences is given. Specifically lineages affecting less than 10 percent of the daily sequences are summed under "Others".

`/explorer/getLastUpdate`: extracts the date of the first and last sequence stored within the database. This GET method is used to inform the user of the time period within which the sequencing data can be analyzed.

`/explorer/getLineagesCharacteristics`: given a list of lineages `LL`, this method returns the list of mutations that are present in at least 50% of the sequences associated with the lineages in `LL`. Notice that these statistics are not computed in real time. Indeed, given the computational effort they require, they are pre-computed and directly extracted from the *Lineages_ Characteristics* table.



/explorer/getLastUpdate

/explorer/getLineagesBreakdown

/explorer/getLineagesCharacteristics

/explorer/getSequenceInfo

Figure 4.9: Swagger Documentation for the Explorer component

4.3.2. Lineage Specific component

The Lineage Specific component is devoted to the computation of statistics related to lineage-dependent analysis. As shown in Figure 4.10, the exposed APIs are:

`/lineage_specific/getAllLineages`: it returns a list of all lineages in the database.

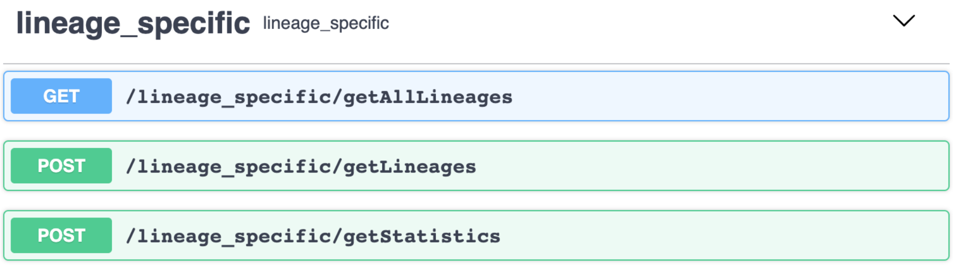
`/lineage_specific/getLineages`: given a date D and/or a location L returns a list of all lineages in the database associated with L and D.

`/lineage_specific/getStatistics`: it is one of the core API. Given a geographic area `A` (either a continent, a country,or a region), a date `D`, and a lineage `L`, it analyzes the trend of mutations appearing in week `[D-6,D]` (i.e., the week ending on day D) and compares it against the previous 3 weeks (i.e., `[D-27,D-21]` , `[D-20,D-14]` and `[D-13,D-7]`). Specifically, first, data concerning the number of sequences collected in each of the 4 weeks, in the context of `A` and `L`, are extracted. Then, considering the week `[D-6,D]`, the mutations affecting at least 0.5% of the sequences associated with `A` and `L` are extracted. Indeed, sequences with lower percentages are ignored as they are irrelevant and would only complicate the interpretation of the data. Let `M` be the mutations of interest, for each of them we compute: the number of times they are present in each of the 4 weeks; and their frequency with respect to the number of sequences associated with `A` and `L`. We also compute the diffusion slope, through linear interpolation of the diffusion (`y=mx + q`) and three meaningful p-values. Their values are computed using a Chi-square test of independence of variables in a contingency table and they are:

- `P-value with mut`: it shows if the population «*with mutation*» is growing differently compared to everything.
- `P-value without mut`: it shows if the population «*with mutation*» is growing

differently compared to everything.

- `P-value comparative`: it shows if the population «*with mutation*» is growing differently compared to the population «*without mutation*».



lineage_specific lineage_specific

GET /lineage_specific/getAllLineages

POST /lineage_specific/getLineages

POST /lineage_specific/getStatistics

Figure 4.10: Swagger Documentation for the Lineage Specific component

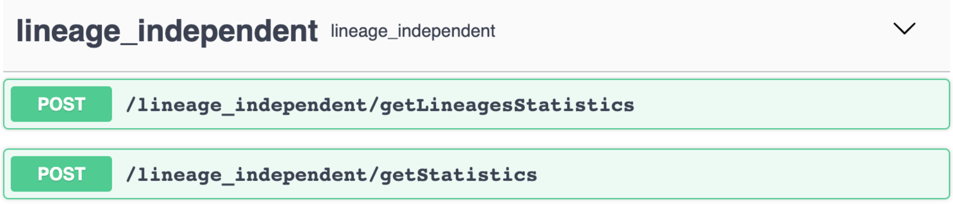
4.3.3. Lineage Independent component

The Lineage Independent component is in charge of computing statistics related to lineage-independent analysis. As depicted in Figure 4.11, the available endpoints here are:

`/lineage_independent/getStatistics`: it is one of the most important API. Given a geographic area `A` (either a continent, a country,or a region), and a date `D` it analyzes the trend of mutations appearing in week `[D-6,D]` (i.e., the week ending on day D) and compares it against the previous 3 weeks (i.e., `[D-27,D-21]` , `[D-20,D-14]` and `[D-13,D-7]`). The first step is to count the number of sequences collected in each of the 4 weeks for the location `A`. Next, considering the week `[D-6,D]`, the mutations affecting at least 0.5% of the sequences associated with `A` are extracted. Let `M` be the mutations of interest, for each of them we compute: the number of times they are present in each of the 4 weeks; and their frequency with respect to the number of sequences in the region `A`. We also compute the diffusion slope, through linear interpolation of the diffusion (`y=mx + q`) and three interesting p-values. The latter are computed using a Chi-square test of independence of variables in a contingency table and they are:

- `P-value with mut`: it shows if the population «*with mutation*» is growing differently compared to everything.
- `P-value without mut`: it shows if the population «*with mutation*» is growing differently compared to everything.

- `P-value comparative`: it shows if the population «*with mutation*» is growing differently compared to the population «*without mutation*».



lineage_independent lineage_independent

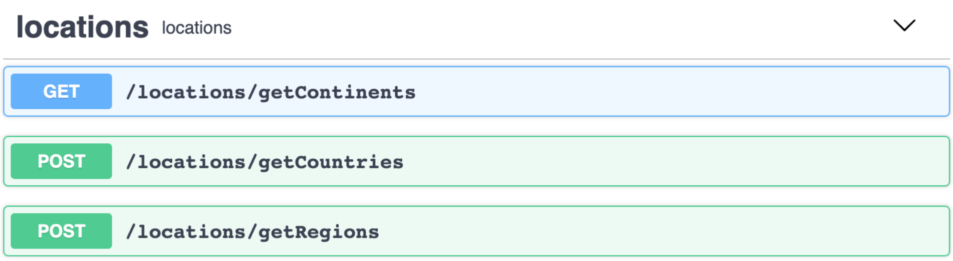
POST /lineage_independent/getLineagesStatistics

POST /lineage_independent/getStatistics

Figure 4.11: Swagger Documentation for the Lineage Independent component

4.3.4. Location component

The main function associated with the Location component is to support the selection of the geographical area of analysis. The exposed APIs are reported in Figure 4.12.



locations locations

GET /locations/getContinents

POST /locations/getCountries

POST /locations/getRegions

Figure 4.12: Swagger Documentation for the Location component

4.4. Frontend

The frontend part was implemented using Javascript, specifically the **Vue.js** framework [27]. As the primary goal of Variant Hunter is to provide a *flexible* tool able to simplify the identification of emerging variants, the GUI was carefully organized to be *intuitive* and *functionality-oriented*.

The large number of features that the frontend part is supposed to provide have required considerable design effort. As a matter of fact, it is organized into a number of Vue components to promote extensibility and re-usability. The frontend is arguably the most complex part of the application due to the multiple data flows and user interactions to be managed. Among the external libraries employed, it is worth mentioning **Vuetify** and **Plotly**. The former is a comprehensive UI framework built on top of Vue.js that provides the basic building blocks to develop rich and engaging user experiences [28].

While instead, the latter is a high-level declarative charting library, used to produce visualizations concerning trends in the spread of SARS-CoV-2 mutations [21].

4.4.1. UX Design

The two core analyses, namely the lineage-specific and the lineage-independent analysis, are arranged in two separate tabs. They share the same basic structure, and the search parameters are synchronized to speed up the switch from one to the other.

Both consist of two parts: a form at the top which allows to define new analysis; and a section including the list of examinations already performed, shown as a list of expandable/collapsible panels.
This makes it possible to compare the results from different queries instantly, which opens up a world of possibilities.

As illustrated in Figure 4.13, the search definition area incorporates the Dataset Explorer functionality, allowing for the accurate selection of the analysis period, as well as the examination of the lineages breakdown over time.

As shown in Figure 4.14, each result panel includes, at the top, a table listing all the mutations, which provide multiple filtering (by protein, by mutation) and sorting options (rows can be sorted based on values in one or more columns).

Specifically, the mutation filters also offer advanced selection modes: via a popup interface the user can interact either by entering a list of mutations directly, or by selecting one or more lineages of interest from a drop-down menu. In the latter case, the characterizing mutations for that(those) lineage(s) are selected.

In the case of lineage-independent search, an option is available to expand each row (mutation) and view the lineage breakdown .
Whereas for lineage-specific search, it is also possible to directly select all non-characterizing mutations. In addition, each row related to a mutation characterizing the lineage under examination is highlighted.

The lower part of the panel, instead, contains three charts: a heat-map, and two line plots. The user can decide to focus on the ones of interest by acting on the show/hide options.

The graphs offer several interaction options: zoom in/out, mutation selection/deselection, download, and more. A label interface is used to move between different versions of the same graph.

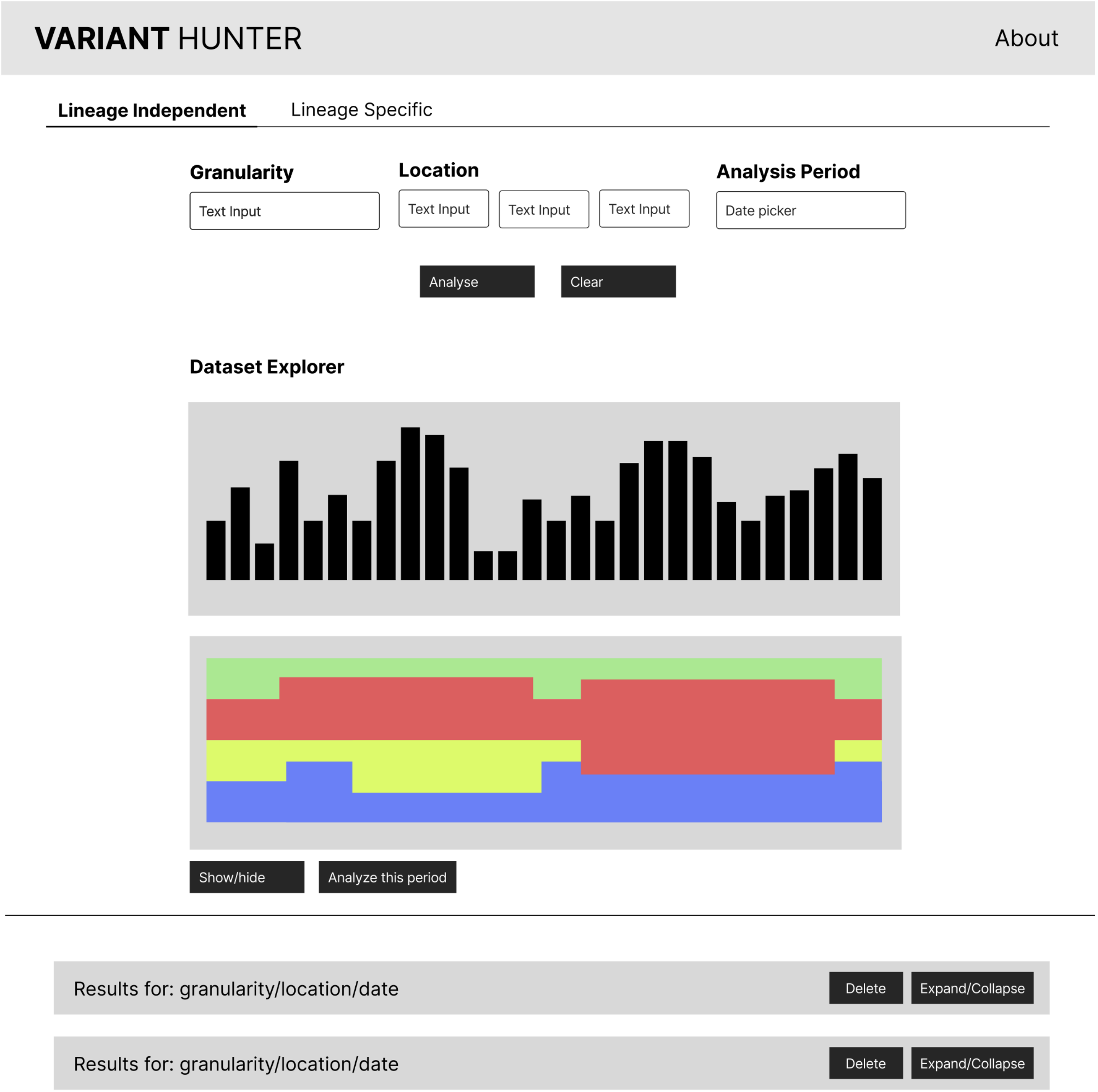


Figure 4.13: Low-fidelity wireframe for the main structure of Variant Hunter

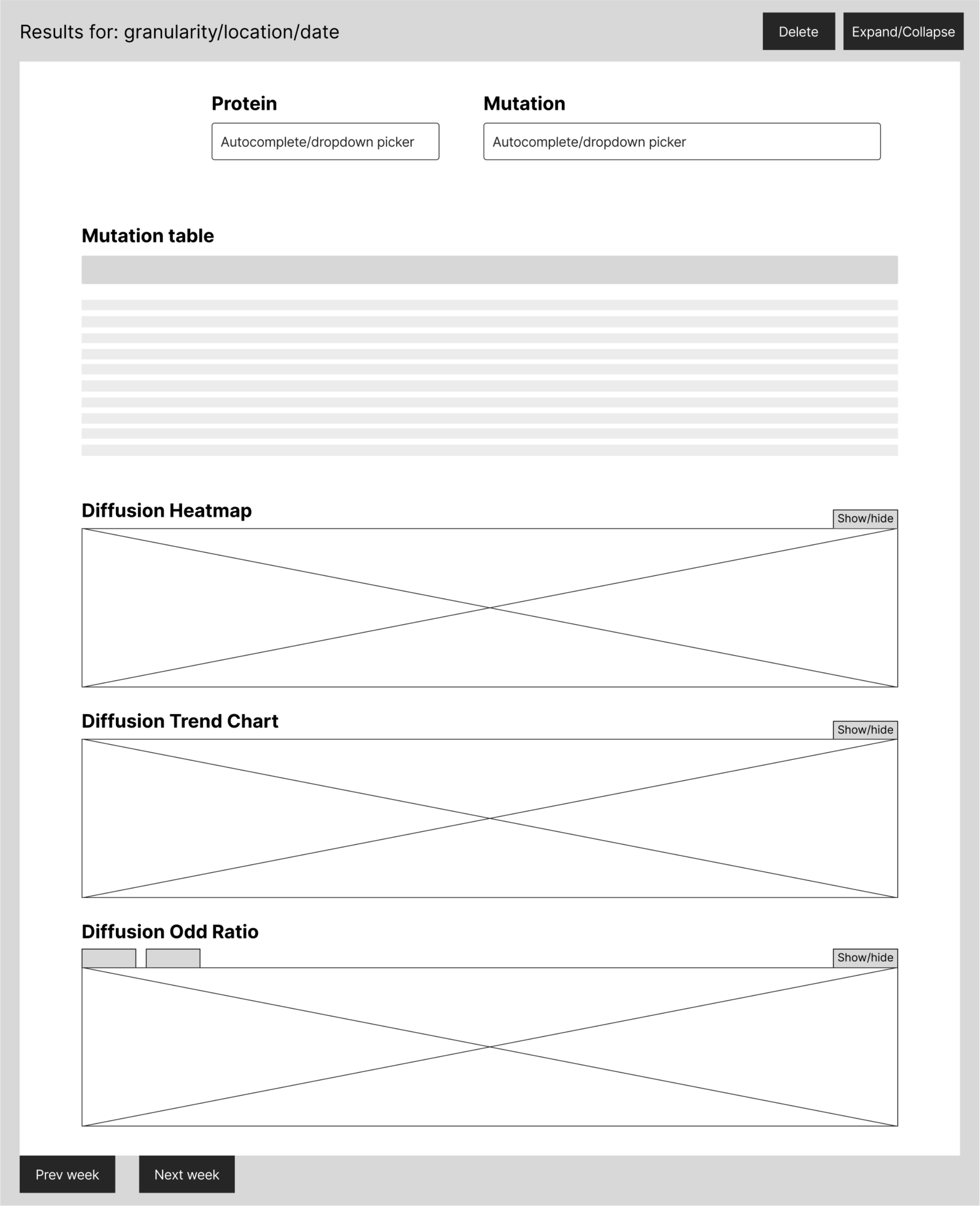


Figure 4.14: Low-fidelity wireframe for the result section of Variant Hunter

By default, all graphs are preset to show the trends of the 5 mutations with the highest growth rate and the 5 mutations with the highest decrease rate. In other words, the 5 mutations with the highest slope value and the 5 with the lowest value are automatically selected.

However, the user has the option of visualizing in the graphs the trends of the mutations he or she wishes to analyze. To do this, he or she simply has to manually select the corresponding rows in the table at the top.

The user also has the option of selecting all the mutations in the table or all the mutations that meet specific filtering criteria (if defined) via an appropriate checkbox in the heading of the table.

Finally, there are two shortcuts to generate a new analysis shifted one week forward or backward, respectively. To further enhance the user experience, new searches generated through these buttons preserve filtering and selection settings.

User interaction follows a linear schema and is supported by plenty of help/explanation labels.

As an example, the interpretation of the plots is aided by special labels that appear on the points of the graph where the user place the mouse.
These enable correct reading of the data and make it possible to display the precise values of the statistics, as well as the number of supporting sequences.

In addition, to allow quick identification of patterns that are not sufficiently supported by the data (i.e., those supported by 5 or fewer sequences), different types of markers are used: a cross indicates that the number of sequences collected in that week is poor (5 or fewer samples); a solid dot, on the other hand, indicates good support.
This way the user can understand at a glance that he or she should not trust too much the trend for that week.

5 | Deployment

This chapter discusses how the application was deployed to end users. In detail, first the publicly available version of the web application is presented. Next, it is introduced the possibility of installing the system locally via Docker and the advantages this brings.

5.1. Online Version

A first way to deploy the developed software is through a publicly accessible website: `http://gmql.eu/variant_hunter`. This online version of the tool is powered by GenBank data, curated and made available by Nextstrain. This has been possible due to the fact that this data sets are public and free of any policy restricting their distribution and/or reuse.

The database behind this version of Variant Hunter is updated weekly as new releases of the metadata.tsv file are published.

The main goal of this deployment modality is to allow a large number of biologists and virologists to use the tool without forcing them to install the Docker version, presented later on. Indeed, the latter, despite considerable design and optimization efforts, may still require time and computational resources that may not be available. It is important to remember that we are dealing with data from millions of sequences published by hundreds of laboratories distributed around the world.

Nevertheless, the main problem with the online version is related to the fact that, as mentioned in Section 4.1, Nextstrain data are much less numerous than those published by GISAID. However, the usage of the latter is unfortunately strongly restricted.

5.2. Docker Version

To overcome the limitations imposed by the Web version of Variant Hunter, the software has been made available for use through Docker. This way, users are free to use any data set they have, including private or restricted-access ones.

5.2.1. The Docker technology

Docker is a container-based open platform for developing, deploying, and running applications. It enables the creation of a container that includes your application, any binaries or libraries that your application depends on, and any configuration details [9]. Once generated, the container runs under the control of Docker, which in turn runs on top of an operating system.

The advantages of this approach are numerous, including [16]:

Portability : the containerized application perform (almost) exactly in the same way on top of any host operating system;

Performances : containers do not contain an operating system (as virtual machines do). This means that containers have a much smaller footprint and consequently are faster to create and start.

Agility : the portability and the performances offered by containers support the development process, making it more agile and responsive

5.2.2. Configuration options

The Docker version has been equipped with numerous configuration options that make the tool highly flexible.

For example, the user can limit the amount of data to be imported from the metadata file, by only considering specific countries (`LOCATIONS`) or time periods (`START_DATE`, `END_DATE`). This significantly reduces startup and database generation delays, as well as storage space on disk.

There is also an option to export the generated aggregate data (i.e., the generated SQLite database file) or import those already in one's possession (`DB_PATH`). This is particularly useful when one plans to stop the container and wants to restart it later using the same data (otherwise deleted by the Docker Engine). Finally, other options allow one to change the server port (`PORT`) or specify the format of the file (`FILE_TYPE`).

All the parameters are given in Table 5.1

Parameter	Description	Required
`FILE_PATH`	Path to the `.tsv` metadata file. *Example:* `FILE_PATH=/Users/rossi/metadata.tsv`	Required, unless importing an existing database.
`FILE_TYPE`	Type of the `.tsv` metadata file. Currently supported: • GISAID: `FILE_TYPE=GISAID` • Nextstrain: `FILE_TYPE=Nextstrain`	Optional. Default: `GISAID`
`LOCATIONS`	Comma separated list of country names whose sequence data is to be imported in the tool. Use `"All"` to import the entire dataset. *Example:* `LOCATIONS="Italy"` *Example:* `LOCATIONS="Italy,Germany,Iran"`	Optional. Default: `"All"`
`START_DATE`	Start date to be considered when importing data. Only the data in the period `[START_DATE, END_DATE]` will be imported into the tool. *Example:* `START_DATE=2021-12-01`	Optional. Default: `Beginning`
`END_DATE`	End date to be considered when importing data. Only the data in the period `[START_DATE, END_DATE]` will be imported into the tool. *Example:* `START_DATE=2021-12-30`	Optional. Default: `End`
`PORT`	The port to be used by the server. *Example:* `PORT=5001`	Optional. Default: `5000`

`REGENERATE`	Flag to drop the current database (if existing) and regenerate the data from the `.tsv` file.
	Example:
	`FILE_PATH=./metadata.tsv`
	`DB_PATH=./folder_to_overwrite`
	`REGENERATE=true`
		Optional.
		Default: `false`

`DB_PATH`	Path to the `varianthunter.db` database file or location where to generate it. The structure of the `.db` file follows the one defined by SQLite.
	Usage:
	• export the database file
	• use a database file resulting from a previous execution, avoiding its regeneration starting from the `.tsv` file.
	Example:
	`DB_PATH=./save_db_here`
	Example:
	`DB_PATH=./fetch_existing_db_from_here`
		Optional.
		Default:
		`internal path`

Table 5.1: List of configuration parameters for the Docker version of Variant Hunter.

Once the desired configuration has been chosen, the user can execute the container using the commands reported in Table 5.2.

Note that the MacOS instructions use a utility script that periodically adjusts the size of the container. This is necessary if the device in use does not have large amounts of free memory (less than 30GB), because of the different way Docker for Mac handles (releases) memory that is no longer in use.

After the completion of data processing, the full-featured application is available at `http://localhost:<PORT>` from a standard web browser.

Windows/Linux command to run the container
`{parameters_list} docker-compose up` where `{parameters_list}` is a space-separated list of parameters.
MacOS command to run the container
`/bin/zsh ./launcher.sh {parameters_list}` where `{parameters_list}` is a space-separated list of parameters.

Table 5.2: Commands to run the Docker container.

Complete installation instructions are made publicly available via a very detailed step-by-step guide (`http://gmql.eu/variant_hunter/about`.).

5.2.3. Technical perspective

From a technical perspective:

- The environment needed by the Variant Hunter application to run is defined within a `Dockerfile`. More precisely, the latter contains the commands for assembling the software.

- The services that make up the application and the configuration parameter mapping is defined in the `docker-compose.yml` file. This YAML file allows the services to run together in an isolated environment.

6 | Performances and Testing

This chapter discusses the performance of the deployed systems.

First, the optimizations and arrangements made in this regard in the design and development of Variant Hunter are presented. Aspects such as database indexing and optimization for devices with scarce computational or memory resources will be covered.

Next, meaningful numerical results from some conducted tests will be reported.

6.1. Optimizations

In order to build a tool that is powerful but also usable in practice, a number of measures have been implemented to reduce the amount of system resources required and minimize application response time. In the following some of the most important ones are given and briefly described.

6.1.1. Pre-computation

As mentioned in Chapter 4, most results are computed beforehand so that an analysis, which could potentially involve several million of sequences, can be performed in less than a second.

In addition to aggregate sequence and mutation statistics, lineage characterization data are also pre-computed. Such computationally intensive operations would otherwise slow down the research workflow excessively.

6.1.2. Database indexing and others

In order to reduce memory requirements, a temporary database is employed during aggregate generation stage to store data supporting the computation.
This database file is deleted as soon as the information is no longer required by the statistics generation software.

This approach makes it possible to overcome SQLite's internal management and immediately release free disk space for other operations.

For the purpose of speeding up access to the database and thus improving the responsiveness of the application, proper indexes were defined on the tables used by the queries.

In particular, attempts were made to balance the occupancy space, the time overhead required for their creation, and the time benefit brought.

6.1.3. Bulk parsing of the metadata file

Mid/low-end computers are often equipped with small memory.
In order to enable high-performance application setup also on these machines, the adopted parsing strategy for the metadata file is structured into two iterative phases:

- first, sequencing data is accumulated in a batch;
- once 50,000 rows are reached, they are loaded into the database in bulk.

6.1.4. OS-specific optimizations

Although Docker promotes itself as a technology that allows an application to run across all operating systems while exhibiting exactly the same behavior [16], in practice this is not always the case.

Indeed, the way the platform manages, and specifically releases, memory that is no longer used by the container can lead to issues, in MacOS, if large amounts of data are generated and removed in very short periods of time.

This behavior, which is not relevant when the available storage space is abundant, can, on the other hand, cause problems if the free disk is limited (less than 30 GB).

In order to compute the aggregate statistics of millions and millions of sequences, SQLite generates a number of (huge) temporary files that it promptly releases when no longer needed. The problem here is related to the fact that they are not immediately released by Docker.

In conclusion, in order for the application to function properly on MacOS in the presence of scarce resources, a script has been implemented to periodically force Docker to resize the container and thus greatly reduce memory requirements.

6.2. Testing

In this section we report the results of some of the numerous tests performed on the final application. With the goal of delivering a high-quality application, several types of assessments were conducted. In detail:

- **User Acceptance Testing** (or end-user testing): it consists of a phase of software development in which the application is tested in the real world, by its intended audience.
  The goal is to ensure that the software can handle real-world tasks and perform according to development specifications and user expectations.
- **Automated testing**: special tools were used to verify that all good development and usability practices were met.
- **Targeted performance and memory tests**: additional tests were conducted to assess the consumption of resources, such as time and memory, required for the Docker version to generate the database.

More details on the results of these audits are given below.

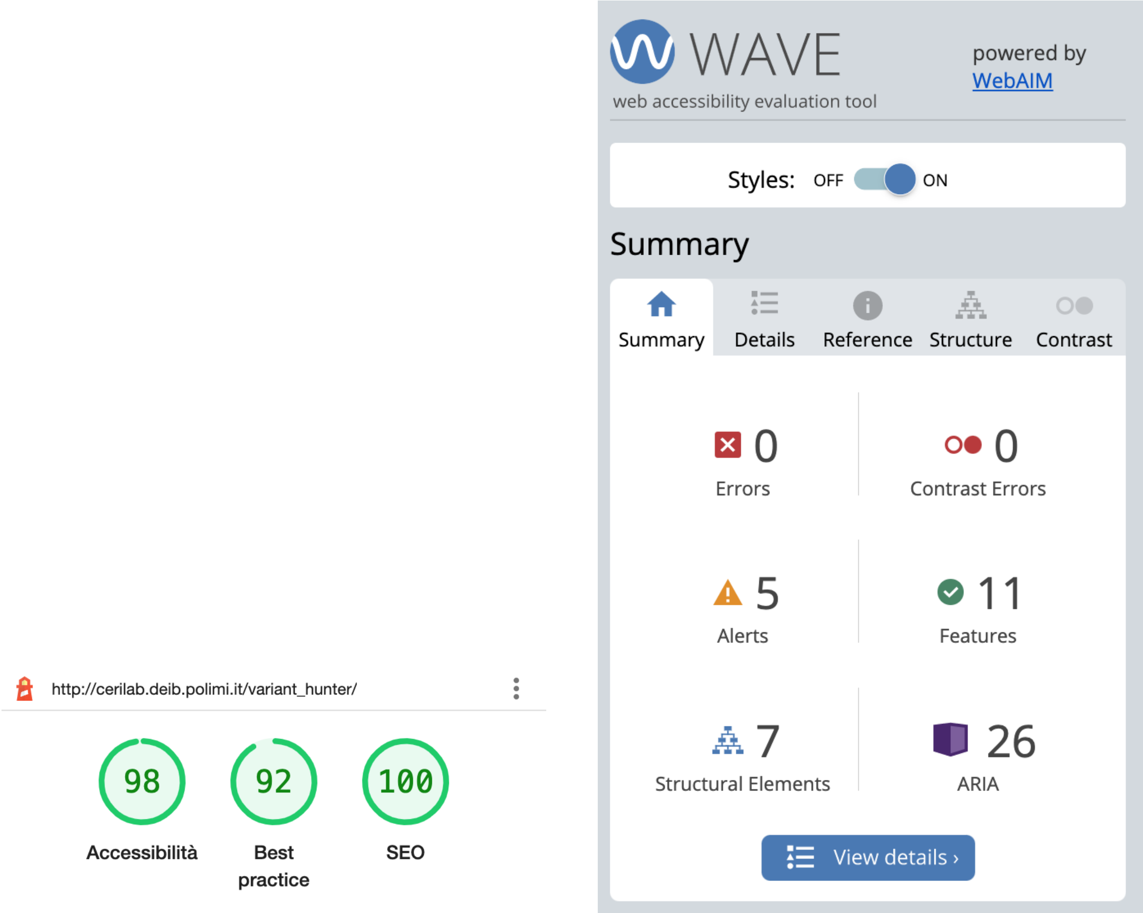
6.2.1. End-user Testing

In order to verify the proper functioning, the application was extensively tested not only internally but also by relying on external research institutions.

Such collaboration has enabled the gathering of direct feedback from what are considered, to all intents and purposes, the target users of the system. This has made it possible to refine and improve the application in order to meet the specific needs of the end users.

6.2.2. Usability, Accessibility and Best Practices

Additional usability and optimization tests were also conducted to identify the presence of critical issues in the application. Specifically, these tests were primarily performed using Google Lighthouse, an open-source, automated tool for measuring the quality of web pages, which tests performance, accessibility, and search engine optimization [12]. Another analysis and verification tool employed is WAVE, «*a suite of evaluation tools that helps authors make their web content more accessible to people with disabilities.*» «*WAVE can identify many accessibility and Web Content Accessibility Guideline (WCAG) errors, but it also facilitates human evaluation of web content*» [29]. As shown in Figure 6.1, Variant Hunter raised no critical alerts from either tool, promoting the application with flying colors.



(a) Google Lighthouse Report.

(b) WAVE Report.

Figure 6.1: Reports from Google Lighthouse and WAVE for `http://gmql.eu/variant_hunter`

6.2.3. Memory and Performances

In the following we provide a summary of other assessments conducted, focusing on the performance that the system was able to deliver.

Table 6.1 and Table 6.2 show the results of generic memory and performance tests on a typical machine available to the target user. The reports reveal how, even in the presence of tens of millions of sequences, the system still manages to provide acceptable performance and storage occupancy.

Machine type	Standard mid-level machine
Machine specs	MacOS, 4 cores, 16GB RAM
Number of sequences	10M sequences (size of GISAID metadata.tsv as of April 2022)
Results	This required ∼30GB (during processing) and ∼2 hrs to startup. In this case, after the database generation, the Docker container required ∼8GB.

Table 6.1: Results of generic tests conducted on the Docker version of Variant Hunter with a mid-end computer, involving 10M sequences.

Machine type	Standard mid-level machine
Machine specs	MacOS, 4 cores, 16GB RAM
Number of sequences	1M sequences (randomly selected from the GISAID metadata.tsv as of April 2022)
Results	This required ∼5GB (during processing) and less then 30 minutes to generate the aggregated database necessary to start the application.

Table 6.2: Results of generic tests conducted on the Docker version of Variant Hunter with a mid-end computer, involving 1M sequences.

Moreover, Table 6.3 summarizes performance and memory statistics for typical use cases, in which the user only imports portions of the metadata file, through the usage of the configuration options explained in Section 5.2.2.
Similar reports are also reported in Table 6.4 and 6.5.

The results show a resource usage that is minimal, efficient, and in line with the requirements identified in the design phase.

In conclusion, the developed software allows the target user to analyze large volumes of data efficiently and quickly, allowing him/her to focus only on the statistics generated by the application.

Machine type	Low/mid-level machine
Machine specs	MacOS, 2 cores, 8GB RAM
Metadata file	GISAID metadata.tsv as of June 2022 (about 11M sequences)
Configuration	`LOCATIONS="Italy"`
Results	This required ∼180MB of storage space and ∼4 minutes for the full setup. After the processing, the Docker container requires less than 115MB

Table 6.3: Results of performance and memory test conducted on the Docker version of Variant Hunter with location-based data import. Data from GISAID.

Machine type	Low/mid-level machine
Machine specs	MacOS, 2 cores, 8GB RAM
Metadata file	GISAID metadata.tsv as of June 2022 (about 11M sequences)

Configuration	`LOCATIONS="USA,United Kingdom"`
	`START_DATE="01-01-2022"`
	`END_DATE="01-05-2022"`

Results	This required ∼6GB of storage space and ∼50 minutes for the full setup. After the processing, the Docker container required less than 500MB.

Table 6.4: Results of performance and memory test conducted on the Docker version of Variant Hunter with location-based and time-based data import. Data from GISAID.

Machine type	Low/mid-level machine
Machine specs	MacOS, 2 cores, 8GB RAM
Metadata file	Nextstrain metadata.tsv as of June 2022 (about 4.8M sequences)

Configuration	`START_DATE="01-01-2022"`
	`END_DATE="01-05-2022"`

Results	This required ∼4GB of storage space and ∼30 minutes for the full setup. After the processing, the Docker container requires less than 800MB.

Table 6.5: Results of performance and memory test conducted on the Docker version of Variant Hunter with time-based data import. Data from Nextstrain.

7 | Conclusions and Future Developments

7.1. Conclusions

Variant Hunter has proven to be a flexible, user friendly and extremely powerful tool for hunting emerging variants and, more in general, for monitoring the evolution of SARS-CoV-2. The sequencing data analysis software helps people spend more time doing research, and less time configuring and running analysis workflows.

Thanks to the collaboration of several virologists in different parts of the world, the tool was designed to meet the specific needs of the field. Thus, the simple and intuitive design of Variant Hunter is combined with the high flexibility brought about by the large number of features, filters, and queries that can be performed within the application.

While more than tens of millions of genomic sequences of SARS-CoV-2 are available through dedicated repositories, their analysis would generally require a significant amount of manual work and engage a huge number of virologists worldwide. Variant Hunter moves toward automating this work and making it available at fingertips through a public web application.

In addition, research institutions also have the possibility to use Variant Hunter to analyze their own or other private sequencing data by installing the Docker version. In particular, the latter is rich of configuration options that enable the selection of the strictly necessary information only, saving time and computational resources.

The goodness of the work is also confirmed by the fact that the appearance of a large number of variants could have been easily predicted through the tool's features, as demonstrated in detail in sections 3.1.3 and 3.2.3.

7.2. Future Developments

As previously stated, Variant Hunter moves in the direction of automating the manual work performed by researchers to analyze mutations detected in sequencing.

Therefore, the next step in this path is to extend the tool by allowing it to perform meaningful pattern mining in the data fully autonomously. This extension, already under development by the same contributors to the version presented in this document, could then potentially warn of anomalous situations in data trends and thus direct the attention of the scientific community accordingly. The entire task could then be carried out without the direct employment of any virologist and could potentially detect all situations that might normally be missed.

Specifically, the extension would be an alerting system capable of:

- autonomously fetching the most up-to-date data from the web portals that are made available by providers such as Nextstrain or GISAID;
- processing them according to the same approach adopted by the current version of Variant Hunter;
- extracting meaningful patterns from the generated statistics;
- automatically publishing the results of this process, so that they are available to governments and scientists to support the evaluation of possible actions to be taken for the safety of the population.

By automating the entire process, it would be possible to respond in a more than timely fashion to stem the spread of new variants or take social distancing measures when it still makes sense.

For example, the tool could generate reports of the current pandemic situation and alert of the unusual growth of certain mutations in a given region of the world. Once warnings are generated, it would then be useful for the system to be able to interface autonomously in some way with other existing tools used by the scientific community. A possible destination for the alerts would be, for example, the Pango Designation [8] repository, nowadays used by virologists to (manually) report new lineages and allow their integration within the Pangolin [19] schema. Indeed, the Pango nomenclature is being used by researchers and public health agencies worldwide to track the transmission and spread of SARS-CoV-2, including variants of concern. Therefore, keeping it as updated as possible is of paramount importance and this extension of Variant Hunter would potentially be able to support this requirement.

Bibliography

[1] Z. I. AnHai Doan, Alon Halevy. *Principles of Data Integration*. Morgan Kaufmann, 7 2012. ISBN 9780124160446.

[2] S. C. Anna Bernasconi, Pietro Pinoli et al. A review on viral data sources and search systems for perspective mitigation of COVID-19. *Bioinformatics*, 22(2):4121–4123, 2021.

[3] Centers for Disease Control and Prevention. CDC's Role in Tracking Variants, 2021. URL `https://www.cdc.gov/coronavirus/2019-ncov/variants/cdc-role-surveillance.html`.

[4] Centers for Disease Control and Prevention. What is Genomic Surveillance?, 2022. URL `https://www.cdc.gov/coronavirus/2019-ncov/variants/genomic-surveillance.html`.

[5] Centers for Disease Control and Prevention. SARS-CoV-2 Variant Classifications and Definitions, 2022. URL `https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html`.

[6] L. B. Chantal Babb de Villiers et al. PHG Foundation: SARS-CoV-2 Variants. *FIND (the Foundation for Innovative New Diagnostics)*, 2021.

[7] S. Consortium. SQLite. URL `www.sqlite.org`.

[8] CoV-lineages. Pango Designation. URL `https://github.com/cov-lineages/pango-designation`.

[9] Docker. Docker Docs. URL `https://docs.docker.com`.

[10] FileInfo.com. TSV File Extension. URL `https://fileinfo.com/extension/tsv`.

[11] E. C. for Disease Prevention and Control. SARS-CoV-2 - increased circulation of variants of concern and vaccine rollout in the EU/EEA, 14th update. *ECDC*, 2021.

[12] Google. Lighthouse. URL `https://developer.chrome.com/docs/lighthouse/`.

[13] J. Hadfield et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, 2018.

[14] Istituto Superiore di Sanità. Sars-CoV-2, 2021. URL `https://www.issalute.it/index.php/la-salute-dalla-a-alla-z-menu/s/sars-cov-2`.

[15] Istituto Superiore di Sanità. Varianti Virali, 2021. URL `https://www.issalute.it/index.php/la-salute-dalla-a-alla-z-menu/v/varianti-virali`.

[16] Micro Focus. Using Enterprise Test Server with Docker, 2022. URL `https://www.microfocus.com/documentation/enterprise-developer/ed40pu5/ETS-help`.

[17] National Human Genome Research Institute. Ribonucleic Acid (RNA), 2022. URL `https://www.genome.gov/genetics-glossary/RNA-Ribonucleic-Acid`.

[18] Nextstrain. Genomic epidemiology of SARS-CoV-2 with global subsampling. URL `https://nextstrain.org/ncov/open/global`.

[19] Á. O'Toole et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evolution*, 7(2):veab064, 2021.

[20] Pallets. Flask. URL `https://flask.palletsprojects.com/`.

[21] Plotly. Plotly: the front end for ML and data science models. URL `https://plotly.com`.

[22] F. RestPlus. Flask-RESTPlus. URL `https://flask-restplus.readthedocs.io`.

[23] R. Sanjuán and P. Domingo-Calap. Mechanisms of viral mutation. *Cellular and Molecular Life Sciences*, 73(23):4433–4448, 2016.

[24] E. W. Sayers et al. GenBank. *Nucleic acids research*, 49(D1):D92–D96, 2021.

[25] C. L. C. Shabir A. Madhi, Vicky Baillie et al. Efficacy of the ChAdOx1 nCoV-19 Covid-19 Vaccine against the B.1.351 Variant. *The New England Journal of Medicine*, 384(20):1885–1898, 2021.

[26] Y. Shu and J. McCauley. GISAID: Global initiative on sharing all influenza data – from vision to reality. *Eurosurveillance*, 22(13), 2017.

[27] Vue.js. Vue.js: the Progressive JavaScript Framework. URL `https://vuejs.org`.

[28] Vuetify. Vuetify: a material design framework for Vue.js. URL `https://vuetifyjs.com/`.

[29] WebAIM - Utah State University. WAVE Web Accessibility Evaluation Tool. URL `https://wave.webaim.org`.

[30] World Health Organization. Global excess deaths associated with COVID-19, 2022. URL `https://www.who.int/data/stories/global-excess-deaths-associated-with-covid-19-january-2020-december-2021`.

[31] World Health Organization. Tracking SARS-CoV-2 variants, 2022. URL `https://www.who.int/activities/tracking-SARS-CoV-2-variants`.

[32] D. S. Xin Luna Dong. *Big Data Integration*. Morgan & Claypool Publishers, 8 2015. ISBN 978-1-62705-224-5.

List of Figures

List of Tables

List of Symbols and Acronyms

ETL

Extraction Transformation Loading.

The ETL process consists of the extraction of data from an initial source, its transformation into a suitable target format, and the loading of the processed data into a target destination (e.g., databases or data warehouses)

TSV

Tab-Separated Values.

A TSV file is a tab-separated values file widely employed to exchange data between databases. It consists of a data table in which each record in the table is on a separate line, and data columns are separated by tabs [10].

Acknowledgements

For the development of this project, I would like to thank my supervisor, Professor Stefano Ceri, who accepted me for my Master's thesis and supported me during all the work.

I would also like to thank all the members of the Geco Team of Politecnico di Milano who collaborated with me throughout the project. In particular, my gratitude goes to my co-supervisor, Dr. Pietro Pinoli, for his helpfulness and kindness during my work. His input and suggestions were essential in achieving the results of this thesis. I also thank Anna Bernasconi, Arif Canakoglu, Matteo Chiara and Erika Ferrandi for their invaluable contributions and collaboration.

Finally, I also thank: Luca Cilibrasi, for his contribution to the first prototype of the system; Shay Fleishon and Valeria Micheli, for their useful feedbacks and comments as end-users of Variant Hunter.

Ringraziamenti

Giunto alla conclusione del mio percorso di studi, mi sento in dovere di dedicare questo spazio alle persone che, con il loro supporto, mi hanno aiutato a raggiungere questo traguardo.

Un ringraziamento speciale ai miei amici e colleghi, in particolare a F. per esserci sempre stato, soprattutto nei momenti di sconforto.
Grazie a R., S. e G., per il vostro prezioso supporto.
Un ringraziamento particolare va anche a F., A. e L., con cui ho condiviso parte di questi anni e che si sono rivelati spesso fondamentali.
Grazie anche a L. e S., compagni di progetto unici che hanno reso questi ultimi anni più piacevoli.

Non posso non menzionare anche i miei genitori, che da sempre mi sostengono nella realizzazione dei miei progetti.

Infine, dedico questa tesi a me stesso, ai miei tanti sacrifici e alla mia tenacia che mi hanno permesso di arrivare fin qui.